DOCUMENT RESUME

ED 177 226 TM 009 816

AUTHOR	Hively, Wells; And Others
TITLE	Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the MINNEMAST Project. CSE Monograph Series in Evaluation, 1.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE	73
NOTE	109p.; Appended item forms may be marginally legible
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Achievement Tests; *Criterion Referenced Tests; *Curriculum Development; *Curriculum Evaluation; *Educational Assessment; Educational Objectives; Educational Testing; Elementary Education; Elementary School Mathematics; Elementary School Science; Evaluation Methods; Evaluation Needs; Experimental Curriculum; Individual Tests; Item Banks; Item Sampling; *Test Construction; Test Interpretation; Test Items
IDENTIFIERS	*Domain Referenced Tests; *Minnesota Mathematics and Science Teaching Project

ABSTRACT

A technology for domain or criterion referenced testing was developed to evaluate the Minnesota Mathematics and Science Teaching Project (MINNEMAST) in grades K-6. The venture was influenced by the low achievement of MINNEMAST students on standardized tests. After instructional objectives were selected and operationally defined, corresponding test items for each course unit were developed. As the collection of items grew, they were arranged in formal tables called item forms. Item forms served two basic purposes: they obviated the need to store individual items, by substituting a set of written rules through which items could be generated when needed; and they enabled the relationships among items to be traced by specifying item characteristics, including stimulus and expected response, and instructions for administering and scoring. The test was actually an ordered set of items administered individually according to a matrix sampling design. Because each child received different item sets, there were as many tests as children sampled. Results were interpreted in terms of patterns of responses to item sets with related stimulus and response characteristics. (Recommendations for curriculum development and evaluation, sample item forms, and a sample interpretation of test results are included.) (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CSE
MONOGRAPH
SERIES
IN
EVALUATION

1

DOMAIN-REFERENCED CURRICULUM EVALUATION: A TECHNICAL HANDBOOK AND A CASE STUDY FROM THE MINNEMAST PROJECT

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
M. Young

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

CENTER FOR THE STUDY OF EVALUATION
UNIVERSITY OF CALIFORNIA • LOS ANGELES

TM009 816

CSE MONOGRAPH SERIES
IN EVALUATION

EDITOR
Marvin C. Alkin

Editorial Reviewer for Volume 1
Howard Sullivan

Center for the Study of Evaluation
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

DOMAIN-REFERENCED CURRICULUM EVALUATION: A TECHNICAL HANDBOOK AND A CASE STUDY FROM THE MINNEMAST PROJECT

Wells Hively, Graham Maxwell, George Rabehl,
Donald Sension, and Stephen Lundin

Center for the Study of Evaluation
University of California, Los Angeles, 1973

CSE MONOGRAPH SERIES IN EVALUATION

VOLUME

1. Domain-Referenced Curriculum Evaluation: A Technical Handbook And a Case Study from the MINNEMAST Project
   Wells Hively, Graham Maxwell, George Rabehl, Donald Sension, and Stephen Lundin

© 1973 by The Regents of the University of California.

In 1968, AERA began its Monograph Series on Curriculum Evaluation. The monographs provided a necessary service to the educational community by presenting a series of articles which contributed significantly to the field of curriculum evaluation. It was a service from which all of us could profit, whether we were concerned with theory, practice, or research and development activities. Now the AERA series, apparently, has been concluded, and those of us who relied upon the monographs for a continuing information service in evaluation recognize the loss: the kinds of goals realized by the AERA series and the service and information it provided should continue to be met. Since we at the Center for the Study of Evaluation are concerned exclusively with research and development in evaluation, we have assumed the responsibility of meeting the need arising from the termination of the AERA series, and have begun to publish the CSE Monograph Series in Evaluation. We will continue the spirit and intent of the AERA series by exposing the educational community to many diverse facets of evaluation.

Each monograph will consist of one or more articles treating a current and relevant issue in educational evaluation. The series will cover the broad spectrum of evaluation activities and issues—theoretical and practical—comprising evaluation, by presenting the work of experts in the field who represent a wide range of educational philosophies and who subscribe to a variety of evaluation frameworks and models. The CSE series will therefore constitute a forum for evaluation and evaluators, provide an opportunity to present important contributions to the field, and continue to meet the high standards and purposes that were accomplished by the AERA series.

This volume is the first in the CSE Monograph Series in Evaluation. We take great pleasure in introducing the series with the work of Wells Hively and his associates on the topic of domain-referenced curriculum evaluation. Although much has been said and heard about the subject, the history of its inception and development and its relevance for evaluation are here presented for the first time. We believe that the potential of domain-referenced testing for educational evaluation is high; we look forward to producing an evaluation monograph series that will realize an equally high potential.

Marvin C. Alkin
Director, CSE

TABLE OF CONTENTS

Foreword xi

Preface xiii

Introduction 1

Chapter I: The MINNEMAST Curriculum Project 3

- General Features
- Aims of the Project
- Instructional Techniques
- Production
- Implementation
- Research
- Evaluation

Chapter II: Domain-Referenced Curriculum Evaluation in the MINNEMAST Project 11

General Rationale 11

- Selecting Instructional Outcomes to Evaluate
- Assessing Knowledge and Ability
- Defining Curriculum Objectives Operationally
- Domain-Referenced Achievement Testing

Developing Test-Item Domains 16

- Choice of Curriculum Segments to Analyze
- The Analytical Dialogue
- Pitfalls of Self-Evaluation
- Procedures for Developing Domains
- The Point-of-Entry Problem

Composing Item Forms 23

- Item Forms
- Item-Form Cells
- The Item-Form Shell
- Description of Materials
- Replacement Scheme and Replacement Sets
- Stimulus and Response Characteristics
- Cell Matrix

Identification of the Item Form
Scoring Specifications

Experimental Design 33

Choosing an Experimental Group
No Control Groups
Post-Testing Only

Test Construction 37

Matrix-Sampling Design
Assignment Sets
The Assignment Plan
Sample Size and Reliability
Limiting Number of Items per Child
Randomizing Order Effects
Practical Procedures in Test Construction

Test Administration 45

Interpreting Results of Testing 47

Categorizing Item Responses
Statistical Inference
Standards of Performance
Interpreting the Results and Revising the Curriculum

A Note on the Cost of the Activity 49

Contributions of the MINNEMAST Project 50

Chapter III: Considerations in the Design of Future Systems of Curriculum Construction and Evaluation 51

Factors that Influence the Effectiveness of Curriculum Development Projects 51

Need for Clear, Cumulative Overall Strategy
Turnover of Staff
Drag Exerted by Production on Evaluation and Revision
Empirical vs. Artistic Approaches to Teaching
Loose Connection between Written Curriculum and Classroom Activities

An Alternative Strategy of Curriculum Development 55

Focus on Specific School Systems
Direction by R and D Specialists, Not Subject Matter Experts

Provision for Systematic Definition and Refinement of Domain-Referenced Objectives
Local Assessment to Provide Performance Baselines
Experimental Design of Instruction
Practicability

Appendices

1. List of Final Units of MINNEMAST Curriculum 62
2. List of Texts, Articles, and Research Reports Produced by Staff Members of the MINNEMAST Project 63
3. Examples of Item Forms Developed by Members of the Evaluation Staff of the MINNEMAST Project 66
4. Examples of Sections of Reports on Selected MINNEMAST Units 76
5. An Example of a General Summary of the Results of Testing a MINNEMAST Unit 87

References 95

FOREWORD

For almost five years now a number of individuals working in measurement, development, and evaluation have been referring to the pioneering measurement activities of Wells Hively and his associates because of the seminal nature of that work. Faced with a host of complex problems stemming from our attempts to improve the quality of educational measurement approaches, the efforts of Hively and his associates to pin down a critical dimension of test item construction gave many of us reason to hope. Yet, until recently it was little more than a hope. For although we could refer people to the work being carried on at the University of Minnesota, there was little published material which offered the kinds of research descriptions or procedural guidelines we all needed. I was elated, therefore, to learn during the summer of 1971 that Messrs. Hively, Maxwell, Rabehl, Sension, and Lundin had finally put together a definitive statement regarding their work. Professor Hively spent that summer in Los Angeles and we shared many conversations regarding domain-referenced testing as he concluded his editing of the manuscript.

I learned much during that summer. For example, although I had viewed the development of domain-referenced achievement testing as a vehicle for improving the congruence between measurement items and the objectives they were designed to assess, I found that the architects of that system more frequently used the approach as a heuristic technique to explore the relationships among behaviors within a given domain. In either event, the approach has much to commend it and this CSE Monograph, the first in a new monograph series on the topic of evaluation, has finally put together in Chapter 2 the details of constructing domain-referenced tests. The mere fact of its availability should stimulate considerable attention to the topic among a variety of educational specialists, and should further the analysis of the efficacy of this type of approach to test construction.

Chapter 1 of the monograph details the MINNEMAST curriculum project itself and provides a useful historical perspective for the domain-referenced measurement system outlined in Chapter 2 and the subsequent analysis of curriculum development presented in Chapter 3. For me, the Chapter 3 analysis was a surprising dividend of this monograph and a bit of frosting on a cake whose quality was already more than acceptable. For those involved in instructional development and formative evaluation, Chapter 3 offers an insightful analysis of the practical problems encountered by those who would engage in large scale curriculum development. A number of thought-provoking recommendations for future strategies are presented.

It is hard to tell whether this initial CSE Monograph by Hively and his colleagues will be best remembered for the description and analysis of the curriculum development project presented in Chapters 1 and 3 or for the effort to describe the item form approach as presented in Chapter 2. Whether we follow their suggestions that item forms be used as a technique for "teasing out and discovering implicit goals" or whether we view it as a powerful way of operationally defining prespecified goals is not clear. Researchers, developers, and evaluators should profit from either.

Professor Hively and his colleagues have assembled a rich document for us to consider. The Center for the Study of Evaluation should be commended for making this fine piece of work available.

W. James Popham
Los Angeles, California

PREFACE

This monograph represents five years of cooperative work in which credit for specific ideas has been more-or-less completely obscured. The entire staff of the MINNEMAST Project contributed in ways impossible to enumerate. Members of the Research and Evaluation Group whose efforts directly contributed to the technology of domain-referenced curriculum evaluation included Paul Johnson and Frank Murray (staff members in the Department of Educational Psychology, University of Minnesota); Eugene Lenarz, Stephen Lundin, Graham Maxwell, Bruce Mussell, George Rabehl, and Donald Sension (research fellows); Luci Johnson, Anne O'Rourke, Ronald Priebe, Ray Sanborgh, Robert Thorndike, Ronald Thurner, David Warren, and Marilyn Walstrom (research assistants); and Timothy Beach (computer programmer).

The five people listed as authors participated most heavily in the discussions which led to preliminary drafting of the manuscript in the summer of 1970. Graham Maxwell produced the all-important first draft. I brought the manuscript to completion with the editorial assistance of Graham Maxwell, George Rabehl, and Donald Sension. Final editing by Ann Dell Duncan and James Burry was enormously helpful.

An opportunity to discuss the work with members of the staff of the Instructional Objectives Exchange and the UCLA Center for the Study of Evaluation provided invaluable editorial help. The assistance of W. James Popham and Marvin Alkin in making that possible is gratefully acknowledged.

The work has been supported by grants to the University of Minnesota School Mathematics and Science Teaching Project and to the University of Minnesota Center for Research in Human Learning from the National Science Foundation, the National Institute of Child Health and Human Development, and the Graduate School of the University of Minnesota.

Wells Hively
Los Angeles, California

INTRODUCTION

1

This monograph has three interwoven themes. It contains: (1) a technical handbook for the application of domain-referenced test theory to educational evaluation: (2) a case history of the curriculum project in which the technology was developed; and (3) an essay on strategies of educational development and evaluation. By interweaving the three themes, we hope to illuminate some of the technical and strategic problems of educational development from angles that are hard to represent in shorter theoretical articles or technical reports.

The exposition is as follows: Chapter 1 gives a brief summary of the history and mode of operation of the Minnesota Mathematics and Science Teaching Project (MINNEMAST). Chapter II treats the rationale that underlies domain-referenced achievement testing and the technology through which it was applied to evaluation of the MINNEMAST Curriculum. Chapter III examines organizational factors that have in the past limited the effectiveness of empirical, formative evaluation in many large-scale curriculum projects and suggests an alternative strategy for the future.

Readers mainly interested in evaluation technology should concentrate on Chapter II. Those whose main interest is in strategies of educational development and evaluation should concentrate on Chapter III, skimming Chapters I and II to put meat on the skeleton of generalities encountered there. Those whose main interest is in the history of the evaluative effort in the MINNEMAST Project should read Chapters I and III, skimming Chapter II.

To escape from superficiality, an adequate exposition of domain-referenced curriculum evaluation required a large fund of examples. These have been collected into several appendices. Our story is told in the text, but the case is "made" by the appendices.

Two particularly important cases can only be made by examining the appendices and are worth emphasizing here. First, that the "item form" approach is not simply a way to operationally define pre-specified goals but is much more fundamentally a way of teasing out and discovering implicit goals. Second, that the power of the "item form" approach lies in the discovery of patterns of student performance that pinpoint the dimensions along which success (or failure) generalizes.

Chapter I

THE MINNEMAST CURRICULUM PROJECT

GENERAL FEATURES

The MINNEMAST Project was a unique and imaginative effort to pioneer the development of a coordinated modern mathematics and science curriculum for the elementary school. It came into existence in August, 1962 under a grant from the Course Content Improvement Section of the National Science Foundation. Founder and first director of the project was Paul C. Rosenbloom, professor of mathematics and director of the Minnesota School Mathematics (MINNEMATH) Center (a center established by Rosenbloom in 1958 for the study of mathematics teaching and administered through the Institute of Technology of the University of Minnesota).

From the beginning, under Rosenbloom's leadership the experimental nature of the curriculum and, therefore, the necessity of establishing its viability through a program of research and evaluation was recognized. A network of trial centers was established across the country to conduct experimental classes and to provide coordinated inservice training for teachers.

When Rosenbloom left Minnesota and joined the staff of Columbia University Teachers College in September 1965, James Werntz, professor of physics, assumed leadership of both the MINNEMAST Project and the MINNEMATH Center. The project continued its work until October 1970.

The project's final major product was a coordinated mathematics and science curriculum for the kindergarten and grades 1, 2, and 3. A listing of the final units for each grade level is given in Appendix 1.

Work on the MINNEMAST Project also resulted in numerous books, articles, and reports by staff members. A list of those produced with the help of MINNEMAST funding is given in Appendix 2. Included are books and articles that provide the teacher with background material and suggestions for the teaching of elementary school mathematics and science, as well as research reports that deal with various aspects of children's learning of mathematics and science.

AIMS OF THE PROJECT

The MINNEMAST Project was devoted to the preparation of a coordinated and sequential mathematics and science curriculum for grades K–6.

Coordination implied concentrating on the interrelations of the two subjects to facilitate transfer of knowledge and skills from one subject to the other and to exploit the possibilities of mutual support for each subject in the teaching strategies.

> In our own Minnesota Mathematics and Science Teaching Project we are aiming to produce a coordinated science and mathematics curriculum for grades K–6.
>
> We have some materials on light for grade 2 which dovetail very nicely with our mathematics program. We have experiments to show that light travels in straight lines in air. On this basis, the children can investigate sticks and their shadows. When they wish to calculate the height of a tree from the length of its shadow, they use similar triangles and have something important to do with multiplication.
>
> Our unit on density, now written for fourth grade, fits in perfectly with our mathematics unit on volume for second grade and our work on coordinates for third grade. If the science group can adapt the measurement of weight to be used in the second grade, then the children could measure weights and volumes of several pieces of iron, and plot weight versus volume on graph paper. They will find that a straight line fits the experimental points very well. (Rosenbloom, no date (a), p. 3.)

The term sequential suggested establishment of a hierarchy in which learning at one level was dependent on learning at previous levels. For example, one of the original guiding principals was that the child learn the operations and relations for each subset of the real number system in such a way that they would work consistently for the whole real number system and could be applied to each succeeding subset of real numbers that the child discovered.

> We want to teach addition in such a way that it will work with all the numbers a child will ever need in school. We want to teach multiplication in such a way that it will work for all the numbers a child will need in school. In our program what happens is that the child learns interpretations of the operations and relations that will work all the way through school, but he learns names for more and more numbers. First he knows only names for the non-negative integers, the counting numbers; later on he will learn names for the rational numbers, the fractions; still later on he will learn names for other numbers in the entire system. As he goes through school, he learns names for more points on a line, but he can apply the ideas that he has at the start to all the new numbers as he goes through school.
>
> What does this imply in our teaching strategy and our curriculum-building strategy? It means that from the start in our program, there is an emphasis on continuous quantity as well as discrete quantity. We emphasize measuring, as well as counting, from the very beginning and we try to teach the children the ideas of the operations—addition, subtraction, multiplication, division—in such a way that they will work for continuous quantity as well as for discrete quantity. We do not neglect the other interpretations. In other words, it's important for the child to know that you can interpret '3 + 4' in terms of counting as one application of the operation of addition.
>
> But we want the basic idea of addition to be: Lay off a distance a,

> then lay off a distance b, starting from 0, and reading off here what (a + b) is. This will work for the entire system. Of course, as the children learned names of points on this side of the line they will find they will sometimes have to lay off the distances to the right, and sometimes the distances to the left; but if they have this idea of addition then they can add 3½ and 4¾ and –2, and so on, just as easily as they can add 2 and 3 as soon as they learn the names for these numbers. (Rosenbloom, no date (b), p. 5.)

Content of the curriculum emphasized two main mathematical structures—the real number system, and the geometry of space—and the main investigative strategies of science—observation, measurement, experimentation, description, generalization, and deduction (Rosenbloom, 1964).

INSTRUCTIONAL TECHNIQUES

For purposes of development and instruction, the MINNEMAST Curriculum was divided into discrete units. Each unit dealt with a collection of related topics under a single title. It was intended that each unit occupy about four weeks of instruction assuming that two class periods would be devoted to it each day.

Each unit consisted essentially of a teacher handbook for a sequence of lessons. The handbook contained general statements about goals, explanatory background material on the subject matter, and lists of materials needed for the lessons. Recommended teaching strategies made use of class question-and-answer sessions, teacher demonstrations, stories, group experimentation and observation, and individual exercises. In 1965, Dr. Zoltan Dienes was a member of the project staff, and many of the subsequent units reflected his philosophy regarding the use of games in instructional activities.

There was considerable emphasis of discovery methods of learning, although the characteristics of this teaching strategy were not formalized. Subject-matter threads or topics were dealt with in the manner of a spiral curriculum in which topics introduced at one level of the curriculum were reviewed and expanded at later levels.

In the teaching strategies, rote memorization was de-emphasized. It was hoped that learning of essential facts and techniques could occur without rigorous drill if two conditions were satisfied by the curriculum. First, children should be interested in and excited by the subject matter so that they would want to go on learning more about it; and second, repetition of factual knowledge of technical skill should be incorporated in intrinsically interesting games and experiments.

PRODUCTION

Development of the units was the responsibility of writing teams, each under the supervision of a chairman. The writing teams consisted of subject-matter experts, teachers, psychologists, editors, artists, and consultants.

Following the pattern adopted by such projects as the School Mathematics Study Group (SMSG), a summer writing session was held each year to outline the content of the units and consider possible teaching strategies. Special consultants were brought in for these sessions, and summer classes were held in nearby schools so that teachers could informally try out some of the projected teaching strategies. Evaluation of these strategies was done by the teachers themselves and emphasis was placed on the viability of the techniques for class participation rather than on their specific effects on student learning.

During the rest of the year, the writing teams continued the work begun at the summer writing sessions, developing the units into complete teacher handbooks and producing the final printed copy. This was the stage during which editors checked and rewrote the text and artists produced the worksheets, photographs, and drawings. When the units were prepared to the satisfaction of the writers, editors, and artists, they were made available to the cooperating schools for trial implementation. The production schedule was arranged so that units could be tried out in the schools according to the specified curriculum sequence.

The strategy for coordinating mathematics and science was first to develop each stream separately, and then to search for links between them. Accordingly, during the first phase of the project (1963–1965), work was directed toward preparing two separate curricula, mathematics and science, supporting each other wherever overlapping concepts were found. By 1965, the development of the mathematics curriculum had proceeded as far as the third grade, while the science curriculum, started in 1964, had been developed for kindergarten and second grade.

In 1965, work was begun on rewriting the separate mathematics units and science into a single integrated curriculum, and by 1967, the project had produced mathematics units as far as the fifth grade and science units as far as the second grade, while integrated units had been written for the kindergarten and first grade. After that, a new plan for direct preparation of the integrated units was instituted. At its termination in 1970, the project had prepared a single integrated mathematics-science curriculum for kindergarten through third grade.

IMPLEMENTATION

At some twenty universities and colleges around the country, MINNEMATH Trial Centers were established. Each trial center secured the cooperation of a number of elementary schools. Each affiliated school agreed to teach the curriculum as provided, to spend at least 200 minutes per week in instruction, to allocate at least two hours per year of each child's time for testing purposes, and to participate in various inservice training courses and summer workshops. The trial centers provided inservice training for participating teachers, administrative services to ensure efficient distribution of the instructional materials to the affiliated schools, and consultative services. At the height of this effort, some 20,000 children were being taught the curriculum in affiliated schools.

The purpose of the trial centers and affiliated schools was threefold: (1) the colleges could provide the nucleus for preservice and inservice training of teachers in the curriculum; (2) implementation on this scale could solve some of the problems usually encountered in disseminating new instructional techniques by building a basis of support for the project around the nation; and (3) the availability of a large number of schools where the curriculum was being taught could possibly enable a very comprehensive evaluation of the materials.

Funding restrictions imposed by the NSF forced a substantial reduction in the number of trial centers in 1965 and almost total elimination of trial centers by 1968. Thus, the trial centers (with the important exception of the Twin Cities Center, which was organized beginning about 1965) played almost no role in the project during the period when its final product was developed and when most of the domain-referenced testing work was done.

RESEARCH

From the beginning, psychologically-oriented research was considered an important adjunct to the development of the curriculum. This emphasis made the MINNEMAST Project fairly unique among NSF-sponsored curriculum-development projects of the time and was largely the result of Rosenbloom's convictions about the relationship of research to development. In fact, various related research projects were carried out in the MINNEMATH Center before the MINNEMAST Project itself was funded. For example, between 1958 and 1962, grants were made available to Ned Flanders and E. Paul Torrance, then of the University of Minnesota, Department of Educational Psychology, for studies of the relation between student learning and the characteristics of teachers, and to Lydia Muller-Willis, then of the University of Minnesota Institute of Child Development, for a study of the learning of mathematics concepts by children.

Nathan Gottfried, MINNEMAST director of research and evaluation from 1964 until 1966, undertook several research studies that had direct relationships with MINNEMAST Curriculum, concerning aspects of learning about length and number. Frank Murray carried on the tradition of research into Piagetian conservation, Paul Johnson investigated the psychology of subject matter, and Grace Dyrud examined various aspects of the learning of science and mathematics. Although these efforts were for the most part tangential to the concerns of the writers of the curriculum, the development of the domain-referenced evaluation model in MINNEMAST probably would not have taken place had it not been for the atmosphere that encouraged fundamental research and practical development.

EVALUATION

The evaluation activities of the MINNEMAST Project were not formulated within the framework of a comprehensive, self-conscious model for evaluation such as has been recently provided by Scriven (1967), Stake (1967a), Alkin (1969) or Stufflebeam, *et al.* (1971). Rather, the development of these recent models largely paralleled the MINNEMAST work (see Stake, 1967b).

To sketch the history of the evaluation activities in the MINNEMAST Project, we may make use of a distinction drawn by Scriven (1967, p. 53) between "intrinsic" evaluation, concerned with appraisal of the curriculum itself, and "pay-off" evaluation, concerned with the examination of its effects on the child. Another distinction, with which the first may be crossed, can be made between "formal" evaluation, in which objective criteria and systematic procedures are used, and "informal" evaluation, in which they are not. This yields the matrix shown in Figure 1.

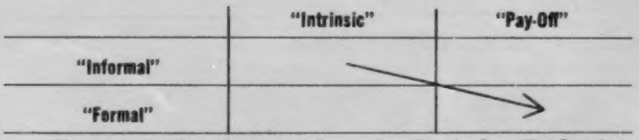


Figure 1: A Classification Scheme for Curriculum Evaluation

In the development of the MINNEMAST Curriculum, all four of the cells of Figure 1 were active to some extent at various times during the project. During the early years "informal intrinsic" evaluations predominated, and procedures were later developed to formalize aspects of the "intrinsic" evaluations. Still later, emphasis shifted to "formal pay-off" evaluations. The general direction of these changes in emphasis is shown by the arrow in Figure 1.

"Intrinsic" evaluation consists essentially of the opinions of specialists, such as subject-matter experts, psychologists, administrators, curriculum supervisors, and teachers. Subject-matter experts (in this case mathematicians and scientists) can examine the curriculum for its logical consistency, correctness, and completeness. Psychologists can assess whether the sequencing is likely to be advantageous to learning, whether the treatment is consistent and developmental, and whether the topics are introduced at satisfactory age levels. Administrators can estimate the feasibility of the program for particular school settings, and predict what modifications might prove necessary in school administration and staffing to implement the program. Curriculum supervisors can predict the difficulties teachers are likely to have with the program and estimate the adequacy of teacher background material and training. Teachers and classroom observers can describe successes and failures encountered in carrying out the classroom activities and make suggestions for improving the stories, pictures, games, vocabulary and reading level, teacher background instructions, materials kits, and teaching strategies. The opinions of all these various specialists were gathered through informal contacts.

Two formal approaches were developed by Nathan Gottfried and used to gather information from teachers during the period 1965 to 1968. One asked all the teachers in the affiliated schools around the nation to fill out a general questionnaire after completing each lesson. The other asked a sample of teachers to keep detailed notes on their teaching of particular units.

Feedback from teachers and from classroom observers naturally included informal pay-off as well as informal intrinsic evaluation. Informal pay-off evaluation typically took the form of anecdotes about children who dramatically succeeded (or failed) in following the classroom activities, and it was also conveyed in estimates of the classroom workability of lessons. At the same time that the staff began to formalize the collection of comments and opinions concerning the curriculum, a first attempt was made to formally test what children were learning. Preliminary drafts of tests were developed and tried out in 1965, but were superseded the following year by the domain-referenced testing activity.

During 1966–67, while the first outlines of the domain-referenced testing procedures were being drawn, an attempt was made to evaluate the mathematics portion of the curriculum using commercially available standardized tests of achievement. Nathan Gottfried and James Ryan (1968) compared the performance of matched samples from experimental and control classes of third- and fourth-grade children selected from schools affiliated with the various MINNEMATH Trial Centers. The experimental group consisted of children who had received only instruction in the MINNEMAST mathematics program from kindergarten onwards. Children in the control group had been exposed to various other curricula than the MINNEMAST mathematics program but came from schools with socioeconomic and educational characteristics similar to the MINNEMAST schools. Matching of individual children was done on the basis of sex and intelligence.

Two kinds of tests were used: (1) two conventional arithmetic achievement tests (Stanford Achievement Test, Intermediate 1, Form W, Test 2, Arithmetic Concepts; and Metropolitan Elementary Arithmetic Test, Form B, Test 1, Arithmetic Computation); (2) a test judged by the MINNEMAST staff to sample outcomes within the scope of both regular and innovative elementary school mathematics programs (School Survey Tests, Third Grade, Educational Opportunities Survey, Part VI). Each child chosen for the study received all three tests. Analysis examined the differences in performance on the tests between experimental and control groups, between sexes, and between children of high and low ability. MINNEMAST pupils did significantly less well on all three tests than did non-MINNEMAST pupils. Gottfried and Ryan called for two kinds of studies in the future: (1) longitudinal comparisons to permit analysis of changes in individual patterns of mathematics acquisition and to examine whether children taught the MINNEMAST Curriculum might, because of the emphasis on structure and process, overtake or surpass the performance of children taught other curricula by the time they reached high school; and (2) replication of their study for other groups at the same grade level to examine the generalizability of their findings.

The next year, 1967–68, Donald Sension carried out a follow-up study on fourth- and fifth-grade classes. A design and analysis similar to Gottfried and Ryan's was used. The main difference was that only one of the MINNEMATH Trial Centers was used. Experimental and control groups were the same as had been drawn from the center for the previous study. Two types of tests were used: (1) three conventional arithmetic achievement tests (Stanford Achievement Test, Intermediate I, Form X, Test 2, Arithmetic Concepts; Stanford Achievement Test, Intermediate I, Form X, Arithmetic Computation; and Stanford Achievement Test, Intermediate II, Form X, Modern Mathematics); and (2) a test compiled by rewriting selected items from the "Arithmetic Concepts" and "Arithmetic Computation" tests in terms appropriate for children being taught the MINNEMAST Curriculum. The results largely replicated those obtained in the preceding year.

The discouraging outcomes of the standardized testing evoked a great deal of controversy. Were the tests really appropriate? Were the main goals of the curriculum adequately represented? Exactly where were the main sources of trouble? The technology of domain-referenced testing was developed in an attempt to untangle such problems.

Chapter II
DOMAIN-REFERENCED CURRICULUM EVALUATION IN THE MINNEMAST PROJECT

GENERAL RATIONALE

Selecting Instructional Outcomes to Evaluate

The basic proposition underlying development of the domain-referenced evaluation model for MINNEMAST was that the crucial question to ask about any program of instruction is, "What effects does it have on the learner?" Given the complexity of school instruction, the amount of information that can be collected about its effects is immense. Therefore, a selection of what effects to study is necessary, since time and resources are always limited.

There are basically two ways to determine what effects to study. One is to ask relevant decision makers what information they would like to have. Another is to examine the expressed objectives for the curriculum. In neither case, however, do the answers come easily. On the one hand, it is possible for a decision maker to ask for information that is inappropriate or incomplete when the logic of his decision making is analyzed. On the other hand, the curriculum writer may state objectives that are too vague to translate into data-gathering procedures. It is necessary, then, for the evaluator to be more than a writer-of-tests and an analyzer-of-data, and to enter into dialogue with the decision makers and the curriculum writers to discover and elucidate their intentions. This is now a fairly frequent theme in the evaluation literature (Scriven, 1967; Stake, 1967a; Alkin, 1969). But in 1966, particularly in NSF projects it was not.[1]

The first step is to locate a sample of representative decision makers. Following the logic of formative evaluation, the decision makers for the MINNEMAST evaluation were taken to be the project staff—writers, consultants, and administrators—whose decisions were directly reflected in revisions of the curriculum goals, recommended teaching methods, and the project organization. This seemed appropriate from the viewpoint of the evaluation staff, which emphasized development of the curriculum itself with little explicit attention to the larger social context in which it was to operate. In retrospect, several other functional, but not necessarily nominal, decision makers should have been systematically consulted from the beginning, e.g., more broadly-sampled representatives of the mathematical and scientific communities, educational policy makers, and representatives of funding agencies.[*]

---

[1] The intellectual atmosphere in which Domain-Referenced Test Theory evolved was to a surprising extent a recapitulation of that surrounding Ralph Tyler's pioneering work. Much of the rationale presented here (and many of the issues discussed in Chapter III) occupies the ground staked out by Tyler (1934) and Smith, Tyler, *et al.*, (1942). In more recent curriculum evaluation theory, the rationale presented here parallels Scriven's description of "mediated evaluation" (Scriven, 1967, p. 55 ff).

The dialogue between the writing staff and the evaluation staff generated a great deal of hard thinking about complex problems. One of the main advantages of such a dialogue is that it forces the evaluation staff to grapple with problems of measuring subtle and long-range goals while at the same time it encourages the writing staff to take advantage of opportunities to gather data on simple, immediate outcomes.

In the MINNEMAST Project, the outcomes identified through this dialogue tended to fall mainly under the headings of knowledge and ability rather than attitude or interest. This was not because of lack of concern with attitudes and interests as outcomes of the curriculum, but rather because measurement of attitudes and interests seemed to require prior definition of related knowledge and abilities.

Attitudes and interests may be thought of as tendencies to engage in behavior that reflects certain knowledge or ability in situations where it is not explicitly required or powerfully evoked. For example, we say a person *knows* how to read if he does it when we specifically order him to or when he is left in a library with little else to do. We say he is *interested* in reading if he does it frequently in situations where other activities are conveniently available. Similarly, we say a person knows how to examine an unfamiliar phenomenon in a systematic, scientific way if he does it when we set one up in the laboratory and ask him to do so. We say he has a scientific attitude if he frequently engages in such behavior outside the laboratory. Although it is difficult to define non-trivial domains for such behavior, in principle, measurement of attitudes and interests can be carried out within the framework of the general Domain-Referenced Testing model to be described below.

Assessing Knowledge and Ability

The traditional method of assessing knowledge and ability has been through the use of standardized tests. It seems natural to use such tests for curriculum evaluation. However, as many have recently noted (c.f. Cronbach, 1963; Stake, 1967; Bormuth, 1970), they have serious disadvantages when applied to evaluation. For example, they are likely to contain items that are unrelated to the content of the specific curriculum being evaluated, and since they are designed to include only those items that discriminate clearly among students, many interesting and important items may be left out because, under "normal" conditions, most children either get them all right or all wrong.

---

[*] See Stake (1967a), Alkin (1969), Stake and Gooler (1971), and Stufflebeam, *et al.*, (1971) for discussions of this problem. In Alkin's terms, the MINNEMAST evaluation activities were concentrated on program improvement to the exclusion of system assessment, program planning, or program implementation.

Standardized tests are constructed to measure hypothetical latent variables such as mathematical ability. Summary scores obtained by adding up the scores from all the items comprising the test are thought of as representing positions along a scalable continuum, and people are judged as possessing more or less of the latent variable according to the magnitude of their summary scores. Comparisons among individuals or groups of individuals are then made by means of the scale. However, from the viewpoint of the curriculum developer, the problem is that two students may know different parts of the subject matter and yet obtain comparable summary scores on the test. From such scores, it is hard for the curriculum developer to tell which parts of the curriculum were effective and which were not.

Similar criticisms may be leveled at tests that evaluation teams construct for themselves if they are built on the same latent-variable model. Such tests may be useful for personnel decisions, but they are of little use in answering specific questions about the detailed effects of particular instruction.[3]

Defining Curriculum Objectives Operationally

Consideration of these and related issues has led various people to propose criterion-referenced testing (c.f. Glaser, 1963, 1971; Popham & Husek, 1969).[4] In criterion-referenced testing, the intention is to clearly specify the behaviors which constitute the intended outcomes of instruction and to design testing procedures which determine which of these a student has attained and which he has not. In general, though, even within this framework, a gap has remained between so-called behavioral objectives and tests that are constructed to measure those objectives.

The most common approach has been to construct prototypical test items that are "keyed" to more generally stated descriptions of the desired behavior (c.f. Mager, 1962; Gagné, 1967; Bloom, 1969; Merwin & Womer, 1969; Lindvall & Cox, 1970). The problem with this approach is that the test constructors cannot claim that the prototype test items exhaustively define the desired behaviors in an operational sense. Given a particular behavioral objective, two independent test constructors often do not come up with the same prototype items. Similarly, given data from the administration of a particular prototype item, independent evaluators often argue about their generalizability to the overall behavioral objective.

---

[3] Steps to the solution of this norm-referenced testing problem have been prescriptively supplied in the form of what might be called "criterion-referenced normed" tests. If published tests and subscales are assigned to meaningful curricular categories on the basis of expert judgments, then adequate measurement in part rests upon the proper selection among those tests that will measure the behaviors in question. This approach has been adopted and successfully utilized in the *CSE Elementary School Test Evaluations* (Hoepfner, *et al.*, 1970), the *CSE-ECRC Preschool/Kindergarten Test Evaluations* (Hoepfner, *et al.*, 1971), the *CSE-HLP Test Evaluations: Tests for Higher-Order Cognitive, Affective, and Interpersonal Skills* (Hoepfner, *et al.*, 1972).

[4] An excellent collection of articles on this topic, with references providing an extensive review of the literature, may be found in Popham (1971a).

A first step toward solving this problem was to try to specify *all* the behaviors which comprise specific pieces of knowledge (Hively, Patterson, & Page, 1968; Osborn, 1968). If this could be done, then the problems of operational definition and of generalizability would seem to be solved. The flavor of this approach as it developed within the MINNEMAST Project is best indicated by quoting from one of the early working papers of the evaluation staff:

> An achievement test is often conceived to be a sample from the set of all possible things that a student could do to demonstrate that he had acquired certain concepts, abilities, or skills. Such a conception is not very useful, unless the domain from which a sample is drawn is well defined. Some people maintain that it is impossible to exhaustively define such a domain of 'criterion behavior' for a non-trivial body of subject matter. Nevertheless, that is how we prepare to approach the direct evaluation of MINNEMAST materials. In its simplest form, the argument for this approach runs as follows.
>
> For the purposes of evaluation and revision we would like to know the proportions of students, in a group exposed to a given unit of instruction, who have acquired various concepts and skills. To find this out we may make up tasks which 'represent' these concepts and operations and ask samples of children to perform these tasks. We may then use the proportions of students who correctly perform the tasks to infer the proportions of students, in the group as a whole, who have learned the concepts and operations. Two kinds of inferences are involved here: a statistical one, in which the proportion of students in the sample who answer an item correctly is taken as an estimate of the proportion of students in the entire MINNEMAST population who could answer *that item* correctly; and an intuitive one, in which we infer from a student's correct response to a particular item that he does in fact 'have' a certain concept and could therefore respond correctly to *other* similar items.
>
> The latter, intuitive inference, is the most troublesome. Still, if we just made up a list of tasks which seemed to incorporate the important concepts and skill at which a unit is driving, we could probably learn a great deal from the proportions of students who answered them correctly (and still more from an examination of the kinds of responses that the students made). We could revise the units on the basis of this information, and measure the effects of the revisions in terms of the proportions of correct answers obtained from a new sample of students who had been exposed to the revised materials.
>
> A critic might object that in doing this we may merely be teaching rote answers to specific items and not the general concepts or skills behind them. The way to find out, of course, is to ask different questions involving the same concepts and skills. We now find ourselves making lists and cross indexes of 'equivalent' items. A logical extension of this activity is to attempt to write the rules which generate *sets* of equivalent items representing clusters of related concepts and

> skills. These are called item forms . . . Items are written as scripts directing the actions of an examiner, with space provided in which to record the responses of a student. Certain elements in the scripts are variable . . . 'Item forms' determine the domains of permissible replacements for these variables. By sampling items from these domains, one can estimate the proportion of students who 'have' the system of concepts and skills represented by the item form as a whole, as well as the proportions who respond correctly to various subcomponents (Hively, 1966).

It required little practical experience in applying these procedures to confirm the literal truth of the claims by such people as Loevinger (1965) that it is not possible to *exhaustively* define "universes" of criterion behavior. Even the simplest concept or skill has so many potential "representative" behaviors that it is impossible to specify them all. Arbitrary limits to the population must be imposed.

Domain-Referenced Achievement Testing

The model that was subsequently proposed recognized that the sets of test items previously referred to as "universes of items" are merely the "nuclei" of hypothetical repertoires of behavior. Such nuclei have been given the more conservative name "domains."

The following excerpt from a final MINNEMAST working paper gives the flavor of the most recent model:

> The basic notion underlying domain-referenced achievement testing is that certain important classes of behavior in the repertoires of experts (or amateurs) can be exhaustively defined in terms of structured sets or *domains* of test items. Testing systems may be *referenced* to these domains in the sense that a testing system consists of rules for sampling items from a domain and administering them to an individual (or sample of individuals from a specified population) in order to obtain estimates of the probability that the individual (or group of individuals) could answer any given item from the domain at a specified moment in time.
>
> Domains of test items are structured and built up through the specification of stimulus and response properties which are thought to be important in shaping the behavior of individuals who are in the process of learning to be experts. These properties may be thought of as stratifying large domains into smaller domains or subsets.
>
> Precise definition of a domain and its subsets makes statistical estimation possible. This provides the foundation for precise diagnosis of the performance of individuals over the domain and its subsets. In addition, clear specification of the properties used to structure the domain makes possible inductive generalization beyond the domain to situations which share those properties. That is, once we have diagnosed a student with respect to a defined domain we may be able to predict his behavior (in a non-statistical, inductive fashion) in natural situations which have some properties in common with the test items within the domain (Hively, 1970).

The foregoing is a rough sketch of the basic notions underlying the domain-referenced testing model. In the following pages, the model will be elucidated in the context of its application in the MINNEMAST Project. In this discussion, emphasis will fall on technical and practical problems, and theoretical issues will be dealt with tangentially. A much more theoretical treatment of the domain-referenced testing model may be found in a thesis by George Rabehl (1971).

DEVELOPING TEST-ITEM DOMAINS

Choice of Curriculum Segments to Analyze

For the MINNEMAST Curriculum, the most logical and convenient segment for analysis was the "unit" since this was treated by the project as an administrative and theoretical building block. Each unit was developed as a distinct set of materials, instructions, and worksheets dealing with a connected set of concepts and skills. A committee of writers and consultants worked together on each unit under the direction of a chairman who assumed major responsibility for the unit and guided it through all the stages of development, production, implementation, and revision. Instruction in each unit was intended to take the teacher about one month. These characteristics made the unit a natural segment for evaluation.

Other segments might have been chosen. Daily lessons might have formed the basis for a more detailed assessment of learning outcomes. However, within the overall framework of the project as it was then conceived, continuous integration of testing into the classroom activities would have been difficult and unpopular.

At the other extreme, the segment chosen for evaluation might have been an entire year of instruction encompassing several units. The advantage of this would have been that the cumulative effects of several units could have been more adequately studied. The drawbacks would have been the difficulty of isolating effects of particular sections of the curriculum and (if one thinks of formative evaluation as a cybernetic process) the enormous feedback delays inherent in annual testing.

The effects of each unit were assessed as soon as instruction in the unit was complete. Long-range or spiraling effects might also have been assessed after instruction in specified units, at the end of each year, at the end of elementary schooling, or even later. However, under the pressure of MINNEMAST production, stress was not laid on such cumulative assessment and the evaluation focused primarily on the immediate effects of instruction in each unit of the curriculum. In future evolution of domain-referenced systems of evaluation, more stress ought to be put on cumulative assessment to pick up such things as interactions of outcomes across units.

The Analytical Dialogue

For the purpose of guiding a domain-referenced curriculum evaluaation, educational objectives as they are usually expressed by writers and teachers suffer from four deficiencies: (1) they often do not specify the actual behaviors that the student should exhibit but refer instead to understandings or appreciations of subject-matter content (e.g., "The objective is to have the children learn that sums of counting numbers can be interpreted as the numbers of elements obtained by combining disjoint sets."); (2) they often specify what the teacher will do rather than what the student will learn (e.g., "The objectives of this unit are to explain the structure of the decimal place-value notation and to provide experience in calculating with denominate quantities."); (3) they are often ambiguous (e.g., "The objective of this unit is to have the children discover and describe the parts a living goldfish uses to move itself from one place to another."); and (4) they are often incomplete (e.g., "The objective of this unit is to develop proper counting procedures.").

These deficiencies arise in part because curriculum writers often do not know in detail what their objectives really are. Instead of beginning with a list of objectives and designing educational experiences to accomplish them, writers and teachers frequently work in the other direction, beginning with ideas for what they feel should be valuable educational experiences and deducing the objectives from the experiences. Thus, important objectives may be implied by the materials a writer develops, but he may be unable to state them clearly or may actually be unaware of their existence. The solution to this problem, attempted in MINNEMAST, was to assign a member of the evaluation staff as a consultant in educational objectives to work with each writing team. The task of the consultant was to arrange a dialogue with the writer(s) in order to arrive at a specification of objectives which both he and they could consider to be satisfactory.

Pitfalls of Self-Evaluation

The close liaison of writer and evaluator raises some questions about objectivity in evaluation. In theory, there would seem to be no reason why the designer of a curriculum could not act as his own evaluator provided that he approach the task *as a scientist*. The scientist works out a hypothesis about some phenomenon that interests him. Since he may have labored hard to produce this hypothesis, he may not be predisposed to believe that it is false. Yet, he suspends his belief and tests his hypothesis experimentally. If the results of the experiment are positive, he will tentatively accept the hypothesis as valid, at least under the conditions operating in the experiment. If, however, the results of the experiment are negative, he will reject the hypothesis as invalid and either begin again to build a new hypothesis or attempt to modify the present hypothesis in such a way as to make it correspond more closely to his experimental observations. The scientific community arranges an elaborate system of social rewards and punishments to keep the individual scientist "honest" in this endeavor.

The curriculum designer is not typically provided with a similar set of social and economic rewards for checking the soundness of his hypotheses. He receives his rewards for producing the curriculum (that is, the "hypotheses"), and he is not rewarded for treating his product as something tentative. Instead, he usually acts as if he believes it is valid and seeks to convince others that he is right. Under such conditions it is reasonable to doubt the evaluative judgments of a curriculum designer concerning his own curriculum.

At the other extreme is the evaluator who has nothing to do with the development of the curriculum and knows nothing about it. It is hardly more likely that he can evaluate the curriculum than the nonscientist can perform the scientist's experimentation. He must, of necessity, find out a great deal about the curriculum he is to evaluate. If, in the process of familiarizing himself with the object of his investigations, he becomes closely involved in the production of some part of it, he may lose some of the objectivity his independence allowed him. How much objectivity he loses probably depends on whether he adopts an active role in which he is critical of and contributory to the curriculum in advance of data collection, or a passive role in which he reserves judgment and limits himself to clarifications of means and intentions. To draw an analogy from chemistry, the difference is between a reactant and a catalyst.

In the MINNEMAST Project, the evaluation staff attempted to act as catalysts in the process of specifying objectives for each unit. That is, they attempted to maintain a detached position with respect to the objectives of the curriculum and sought to clarify those objectives through discussion with the writers.

Procedures for Developing Domains

The task of determining instructional objectives may be thought of as one of moving from vague to specific statements of the properties that the instructional outcomes should possess. An analogy between objectives and architectural blue-prints may be instructive. The blue-print is a definition of the properties that the finished product is to have. It is produced as the result of discussions between the architect (as consultant) and the person for whom the house is to be built. During these discussions, the architect seeks progressively to clarify the intended meaning of the term "house." The blue-print becomes the operational definition of house, and an independent observer is able to check whether the product of the builder's efforts possesses properties that match those specified in the blue-print.

For educational objectives to be operationally defined, the properties of the behavioral outcomes of instruction must be made explicit. General descriptive statements usually contain meanings that remain too subjective and ambiguous. Just as there is less practical value in the word house than the blue-print that defines it, there is less practical value in the phrase understanding place-value notation than the specification of a set of test items (together with a description of their important characteristics) that could be taken to define the understanding. (Although the point has been implied several times in the preceding discussion, it is important to emphasize that the term test item refers to any replicable situation in which specified behavior may be recorded and scored, including non-verbal as well as verbal, paper-and-pencil performance.)

Instead of attempting to describe step-by-step procedures for developing domains of test items, a few predominant strategies will be sketched. For the outlines of a much more thorough, systematic treatment, see Rabehl (1971).

One strategy is to start with prototype items and systematically alter parts of those items to generate sets of equivalent items exemplifying the general concepts or skills supposed to be tapped by the prototype. A preliminary list of prototype items may often be extracted directly from the written curriculum materials. The next task is to determine whether satisfactory performance on each of these prototype items is *really* an anticipated outcome of the instruction and to what extent the list of items may represent a biased subset of some larger set of implicit outcomes. To do this, the list of items may be presented to the writer(s) and he may be asked to make judgments about the acceptability of each item and the completeness of the list. "Is this what you would like students to be able to do? Is there anything else you would like them to be able to do? In what ways might the items on this list be changed and still get at the same general ideas?" This line of questioning results in the production of sets of items that the writer can classify (to his own satisfaction) in terms of general descriptive titles. The resulting pool of items may be small or large depending on the subject matter. It may consist of a simple listing of items or it may be composed by stating sets of rules by which items may be generated. In general, the task is to produce a pool that the evaluator and writer agree upon as representing the central body of knowledge and skill that are the goals of instruction.[5]

---

[5] In MINNEMAST the dialogue was carried on exclusively between evaluators and writers, but it is important to remember that the writer is just one of a number of possible participants or "informants."

The foregoing strategy may be thought of as primarily inductive in nature. Starting from a list of items extracted from the teaching materials, and working through a dialogue with the writer, the evaluator makes inferences about general goals and fills out the set of items accordingly. This strategy may, of course, be balanced by a deductive one in which the dialogue starts from general statements of objectives and the writer is asked to suggest items exemplifying the general goals.

MINNEMAST evaluation staff members varied in the extent to which they made use of the more-or-less standard language of behavioral objectives ("Given A, B, C, the student will do X, Y, Z"). Some found it easier to accept the writers' objectives in their own terms, whether cognitive or behavioral, and to link these to examples at the level of test items. Others attempted to work through an intermediate level of behavioral specification.

A third strategy for filling out a domain involves the formation of hypotheses about sequences or hierarchies of instruction. Given an item, or set of items, representing agreed-upon goals, the evaluator may ask the writer, "What knowledge or skill would the student need before you could teach him this? Are there alternative ways of arriving at the solution or acquiring the skill? Are these things taught in this unit? (Should they be?) Should we devise items to test these abilities, whether or not they are explicit goals of this unit?" Similar questions may be asked, proceeding in the opposite direction: "Given that students *can* do these things, what *else* could you teach them? Are these things taught in this unit? (Should they be?) Should we devise items to test these abilities, whether or not they are explicit goals of this unit?"

This process of tracing out hypothetical connections to prerequisite, or subsequent, knowledge and skill is open ended. It is important, however, to note the considerations that enter into the process of deciding whether to stop or to further expand the domains;

1. Everything that is explicitly taught in the unit probably should be tested. If only terminal behavior is tested it may be hard to trace out the exact sources of failure.
2. Even if prerequisite behavior is not taught in the unit, it probably should be tested (within limits of time and energy) since failure that might otherwise be attributed to the unit may be due to inadequate preparation.
3. It may be worthwhile to include subsequent behavior, beyond the explicit goals of the unit (within the limits of time and energy) to assess generality or transferability of the instruction.
4. Circumstances that set practical limits to the detail with which prerequisite and subsequent behavior may be tested include:
    a. Time and other resources available to writer(s) and evaluator(s).
    b. The size of the student population which, in a matrix-sampling test design determines the amount of ground that can be covered,

given that there are upper limits to the amount of testing to which an individual student may be subjected. The larger the population of students, the greater the range of behavior that may be assessed.

In his dialogue with the writer, following any of the above strategies, the evaluator may either play the role of non-directive questioner ("What items might exemplify this objective? How could these items be changed and still test the same thing? Does this list of items get at all the important objectives of unit? What items might test abilities prerequisite or subsequent to these?"), or he may play a role more analogous to that of a descriptive linguist, working with an informant to discover the structure of an unknown language. That is, the evaluator himself may generate or create permutations of items and try them out on the writer to see whether or not he accepts them. ("Would this item exemplify this objective? If this item were changed in the following way, would it still test the same thing? Would these items be useful additions to the pool for the unit? Would these items test appropriate prerequisite and subsequent abilities?")

Out of their experience in the foregoing dialogue, members of the MINNEMAST evaluation staff acquired a repertoire of what might be called item transformation rules—more-or-less routine ways of altering various characteristics of items to generate new instances that could be presented to the writer for his comments. Rabehl (1971) has taken steps toward working these transformation rules into a classification of behaviors encountered in science, and he has proposed taxonomic principles to be applied within a framework of analysis referred to as Behavioral Systematics. The next generation of work in domain analysis may well draw heavily on Rabehl's theoretical work. However, the MINNEMAST activities represented an intuitive, pre-theoretical stage from which Rabehl's treatment grew.

The following is a rudimentary list of characteristics typically considered by MINNEMAST evaluation staff in producing transformation of items.

1. *Forms of instructions:* instructions may be given in written or spoken form (or a mixture of both) or the situation may be "staged" nonverbally.
2. *Forms of responses:* the student may be asked to produce a written, a spoken, or a motor (i.e., manipulative) response or some combination of the three.
3. *Syntax:* syntactical transformations of the instructions can generate alternative language without changing the response requirements.
4. *Vocabulary:* words used may be technical or nontechnical. Sets of acceptable and unacceptable synonyms may be important to consider.

5. *Sequencing:* the order of separate parts of the instructions can sometimes be permuted.
6. *Supplementary information:* items may be constructed in such a way that, if a child fails to respond, he may be given further specific information. Different classes of items may be created by supplying different kinds and amounts of supplementary information.
7. *Numerical values:* ranges of numbers, lengths, areas, volumes, etc., are very important and have to be specified.
8. *Physical characteristics of materials:* materials that the child must manipulate to arrive at a response usually have physical properties that can be varied systematically to create a range of such materials (e.g., the general physical properties of the class of objects called beam balances can be identified; a child who can operate simple beam balances successfully cannot necessarily operate more sophisticated ones successfully).
9. *Production, choice, or judgment:* the student may be asked to write or construct something that is not already part of the material presented to him; he may be asked to select the correct answer or answers from alternatives given; or he may be asked to judge whether or not a given answer is true or false.

It is important to consider the above kinds of item characteristics carefully because underlying them are implicit theories about performance generalization and instructional transfer. For each transformation performed on an item, one should ask, "How might changes in this characteristic be expected to affect the student's ability to respond? What specific conditions, or ranges of values, are most important? Where can generalization be expected and where will explicit teaching probably be necessary?" The result of such an endeavor is the construction of a theory (perhaps a primitive one initially) about the variables that are likely to affect the student's behavior, the experiences he ought to receive in instruction, and the range of generalization to be expected in his subsequent performance.

In the early stages of this kind of theory making (exemplified by MINNEMAST), it is often hard to tell which item characteristics are trivial and which are important. Such things as whether the instructions are written or spoken may be trivial in some circumstances and extremely important in others. Underlying the domain-referenced testing approach is an assumption that such details of stimulus conditions and response characteristics are more often important than not, and that failure to take them into consideration may account to a great extent for our typical failure to find differences between gross educational treatments. One way to determine if this assumption is true is to carve up some subject-matter domains in detail and study them to see whether such fine-grained conditions do indeed have sizable effects for given groups of children.

The Point-of-Entry Problem

Frequent outcomes of the dialogue between writers and evaluators in MINNEMAST included clarification of latent or unrecognized goals, discovery of new goals, and a kind of restructuring of the subject matter at an epistemological level. These in turn often suggested new teaching procedures or "embodiments." Therefore, when the dialogue took place with respect to a more-or-less completely written unit, the results were often frustrating, since it was often too late to incorporate the new material in the unit. This placed the writer in a difficult position, one in which his creative participation in the dialogue tended to be punished by the discovery of material that he could not immediately use. The more active his entry into the dialogue, the more thoroughly he was likely to expose the shortcomings in his original thinking.

The point-of-entry problem for the evaluator thus turns out to be of great practical importance. If he enters too late, the dialogue is likely to be unproductive. It may also be possible for him to enter too early. Attempts to involve evaluation staff in a preliminary planning of MINNEMAST units sometimes injected a premature note of concreteness that may have inhibited the work of the creative staff. The first-draft stage was found to be a convenient entry point.

COMPOSING ITEM FORMS

Item Forms

As the collections of items representing the objectives of a MINNEMAST Unit grew, they were laid out in formalized schemas called *item forms*. Item forms serve two basic purposes: (1) they obviate the necessity to store individual items by substituting a set of written rules through which items can be generated when needed; and (2) they enable the relationships among items to be traced by giving clear specifications of relevant item characteristics. Thus, there are two major parts to any item form—one that tells how to generate the items and another that describes their salient characteristics.

The MINNEMAST evaluation staff eventually arrived at an item-form format fairly general in application. This format took a long time to develop. Those who examine the file of MINNEMAST item forms will discover that early attempts bear little resemblance in detail to later ones. Throughout the project the item-form format underwent continual evolution. The "final format" should be regarded as an arbitrarily fixed point on a continuum of development. However, any way of writing an item form needs to make provision for the same basic kinds of information.

Figure 2 is an example of an item form. (Several others examples of item forms representing a variety of different types of behavior may be found in Appendix 3.)

ITEM FORM 2.2*

Producing examples of simple and non-simple, open and closed curves.

GENERAL DESCRIPTION

The child is given an example of a simple open, simple closed, non-simple open, or non-simple closed curve and asked to draw several more that are different, but of the same kind.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

Child is given an example of the required type of curve at the beginning. Child produces curves by drawing them.

Distinguishing Among Cells

Type of curve required: (1) simple open, (2) simple closed, (3) non-simple open, (4) non-simple closed. (The last two curve types are not standard topological classifications, but are clearly defined.)

Varying Within Cells

Instances of sample curves presented.

CELL MATRIX

	Script (b)
Simple closed	(1)
Simple open	(2)
Non-simple closed	(3)
Non-simple open	(4)

(Sample curve is drawn from replacement set corresponding to script.)

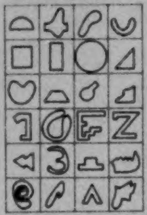
ITEM FORM SHELL

MATERIALS Curve card (a) Response Sheet Pencil	**ITEM FORM: 2.2** CELL: 1 REPLICATION: 1
DIRECTIONS TO E	**SCRIPT**
Don't look at curve card yourself, until you have laid it in front of S.	Here is a (b) simple closed curve. Here is a pencil and paper. Draw another (b) simple closed curve that is different from that one. (Answer #1)
After S finishes each answer, write its number beside it. If you aren't sure whether S is finished, ask him.	Now, draw another (b) simple closed curve that is different. (Answer #2)
In transition to each new question, you can say "um hum" or "O.K." but don't say "good" or otherwise put special emphasis on correct answers.	Now, draw another (b) simple closed curve (Answer #3)
	Now draw another (b) simple closed curve that is different. (Answer #4)
RECORDING Attach response sheet.	

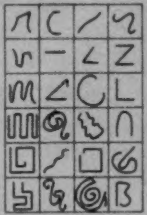
REPLACEMENT SCHEME

Curve Cards (a)	Script (b)
Cell 1: choose from R.S. 2.1.	Cell 1: simple closed
Cell 2: choose from R.S. 2.2.	Cell 2: simple open
Cell 3: choose from R.S. 2.3.	Cell 3: non-simple closed
Cell 4: choose from R.S. 2.4.	Cell 4: non-simple open

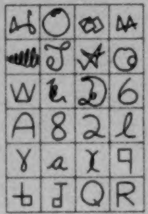
REPLACEMENT SETS



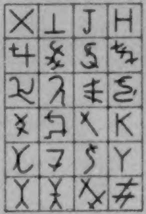
R.S. 2.1.
Simple, closed curves



R.S. 2.2.
Simple, open curves



R.S. 2.3.
Non-simple, closed curves



R.S. 2.4.
Non-simple, open curves

SCORING SPECIFICATIONS

Cell 1 (simple closed): Curve bounds an area, may not have crossing points.

Cell 2 (simple open): Curve does not bound an area, may not have crossing points.

Cell 3 (non-simple closed): Part of curve bounds an area, must have at least one crossing point.

Cell 4 (non-simple open): No part of curve bounds an area, must have at least one crossing point.

Figure 2: Example of an Item Form

---

* Originally developed by Stephen Lundin.

It takes a little practice to read an item form without bogging down in details. First look at the *title* and *general description* at the top of the left-hand panel. That gives a general idea of what the item form is about.

Then look at the *shell* that appears in the prominent box in the center panel. This gives an example of an item as it would be read by an examiner and administered to a student. The shaded parts of the shell denote variables that may be replaced to create other items.

Next, look at the *cell matrix* in the left-hand panel. That gives a general picture of the set of possible replacements for the variables in the shell.

To go into detail about the theoretical characteristics of the replacement structure, look at the *stimulus and response characteristics* in the left-hand panel. To go into detail about the exact mechanics of generating replacements study the panel at the right. What follows now is a detailed discussion of each of the components of an item form.

Item-Form Cells

The smallest classes of items into which a domain is stratified are called Item-Form Cells. An item form may include just one cell or, more usually, several related cells. Cells are grouped into item forms mainly on the basis of convenience and efficiency. Items from several cells may have such similar characteristics that they can be generated from essentially the same set of rules with only minor variations from cell to cell. Thus, grouping these cells together into a single item form removes the need for duplication of some generation rules. In addition, the grouping allows clearer specification of the similarities and differences among the cells.

No precise guidelines presently exist for grouping cells into item forms. It seems best to group together as many cells as possible with the upper limit being determined by the difficulty of writing the generation rules and the difficulty of interpreting the item form. For example, consider the domain of items represented in Item Forms 9.7 and 9.8 (Appendix 3). The items are all part of the domain dealing with the behavior of producing a number satisfying a given relation to a specified number or numbers. In one item form all items ask for a verbal response; in the other, a written response. Otherwise, the two item forms are identical. An attempt was made to cast all of these cells into a single item form, but because the generation rules were difficult to write and understand it was more satisfactory to divide this domain into two separate item forms. Other divisions of the domain could have been made, but division on the basis of the response requirement seemed most convenient and most intelligible.

The Item-Form Shell

The Item-Form Shell contains the common, unvarying components of all items generated by the item form. To produce an item, blank spaces in the shell are filled according to specifications given in the Replacement Scheme. The item then tells an examiner exactly what to give the child and what to say to him. Each item is a set of instructions for the examiner to act upon. These instructions must tell the examiner what materials (if any) to give to the child, when to give them and how to give them, what to say to the child and when to say it, what features of the child's response to observe and how to record them. These instructions are given in four subsections of the shell: Materials—which lists by name the manipulative objects or printed materials needed for presentation to the child; Directions—which describes how the examiner is to proceed in administering the item; Script—which is coordinated with the directions and specifies exactly what the examiner is to say in administering the item; and Recording—which indicates what the experimenter should look for in the child's response and how he should make a record of it. A fifth subsection of the shell identifies the item by code number and allows space for inserting information specifying the child to whom the item should be presented. The total layout allows the examiner to read each item systematically from the top of the page to the bottom.

The layout of the shell is best determined by administrative convenience. The items developed by the MINNEMAST evaluation staff were designed for individual administration by an examiner. Therefore, the filled-in shell had to give clear directions to the examiner concerning what to do and say and to whom. However, a modified shell would be necessary for other forms of test administration such as paper-and-pencil examinations where the child rather than the examiner reads the directions and script and records his own responses.

The convention adopted for indexing replacement variables in the shell is a blank prefixed by a letter referring to the relevant specification within the Replacement Scheme. Replacements can be words, phrases, symbols, numbers, figures, or indices referring to special test objects or materials. They can occur in any part of the shell and even, when there are variables nested within other variables, in the Description of Materials and the Replacement Scheme sections of the item form.

The instructions for recording a child's response should not require an examiner to decide on the correctness of the response. If the response is to be verbal, then the instructions require an examiner to record that response verbatim. If the response is to be manipulative, then the instructions indicate what specific features of the child's actions to observe and how to make a record of them. Where sufficient of these features can be anticipated in advance and where identifying the occurrence or non-occurrence of special manipulations is important, a check-list can sometimes be provided. Even so, the examiner may need to make further written descriptions of the child's response if the check-list proves to be incomplete.

The MINNEMAST item forms did not often supply check-lists but usually required written descriptions of each response. One useful device was the provision of an incomplete diagram that could be filled in to show how the child arranged or used the materials. For example, a stylized diagram of the arms of a beam balance was useful for showing where the plumb-line was resting when the child made his verbal or written response concerning the relation between the weights of two given objects (see Appendix 3, I.F. 16.14). In this case, the diagram provided not only a record of what the child did with the objects but also allowed a check on whether he was given the correct objects by the examiner and whether the balance acted in the anticipated manner. In cases where a written response was required, it was a simple matter to preserve the response on a sheet of paper and identify which item and child produced it.

Description of Materials

The Description of Materials gives a complete description of the test objects that are merely listed by name in the materials section of the shell. To be complete, this description must include all the necessary information and instructions for preparing test objects for specific items. The description should leave the least possible ambiguity about what the test objects are and how they are constructed.

For the MINNEMAST item forms, it was found convenient to label each test object with a Test-Object Number so that each object could be easily cross-referenced if necessary. A complete listing of test objects for each unit was then made in a Test-Object Inventory (a separate document from the item forms). One numbering system made use of a three-part serial number, specifying the curriculum unit, a test-object set, and an element within the set. The convention when referring to a test object from a set of one element was to label it, for example, as T.O. 10.1.1, meaning test object 1 for Unit 10. Entire sets were then referred to, for example, as T.O.S. 12.3.0, meaning test-object set 3 for Unit 12. The first element of this set was T.O.E. 12.3.1. These conventions were found to minimize ambiguity and misunderstandings.

The Test-Object Inventory listed the complete set of test objects for a unit in numerical order. Ideally, the inventory gave complete instructions for the construction of the objects, repeating anything that had already been given in the Description of Materials sections of the item forms. It often happened that complete instructions were difficult or time consuming to write and the actual test objects or their prototypes were maintained in storage and merely referenced by the test-object number and a brief description of salient features.

In a strict sense, the Test-Object Inventory is redundant since it reproduces materials already found in the Description of Materials sections of the item forms. However, there are advantages in this arrangement. First, it is desirable to have a complete specification of the test objects within an item form so that each item form can stand by itself without the need for any supplementary explanation. Second, it is desirable to have an inventory of all test objects to aid in cross-referencing (once an object has been designed, the same test-object number can be used for it at all times), to allow short-cutting in writing the item form (valuable time can be saved by referring a typist to the inventory when an object that is already listed is to be described in an item form), and to provide a complete set of instructions to anyone whose task it may be to assemble the test objects (instead of having to dig through the item forms themselves with the possibility of missing something, the inventory can be used directly for test-object construction).

Replacement Scheme and Replacement Sets

The Replacement Scheme specifies how to choose values or prescriptions for each of the variable parts of the item form. Replacements may be required in any or all of three different places: (1) in the Shell, where variables may be indicated within the Materials, Directions, Script, and Recording subsections; (2) in the Description of Materials, where variation may be indicated within a test object or where choice among test objects may be required; and (3) in the Replacement Scheme itself where variables may be nested within other variables.

The following conventions help to provide uniformity and a minimum of confusion to anyone reading the item forms:

1. Replacement instructions are ordered by variable, each variable being indicated by a letter. An alternative would be to arrange everything by cells. At first this might seem to be a better arrangement since it would probably be easier for the novice to read. But its disadvantage is considerable repetition of information, especially when the number of variables is large. By arranging the replacement instructions by variable, repetition is eliminated. Generation of the items is also facilitated by this arrangement provided that all cells of the item form are being used for the testing.

2. Variables are sometimes given a name to help in identifying where they occur in the item form.

3. Replacement rules are of two kinds. When a replacement is constant for all items within a cell, then it can be specifically identified by means of a verbatim entry enclosed in quotes. When a replacement

is variable within a cell, then it has to be sampled from a Replacement Set according to a sampling directive. Replacement Sets usually consist of words or numerals since they specify what is to be written into the blank spaces of the item form.

MINNEMAST item forms were identified by a two-part numeral in which the first indicated the unit where it was first used and the second part indicated the set number within that unit. Thus, for example, R.S. 16.4 meant Replacement Set 4 Unit 16.

Because Replacement Sets can be used several times within one item form, they are most conveniently collated in a separate section of the item form. Sampling directives can then reference a Replacement Set by number. For the reason that many of the same Replacement Sets are used across many item forms even from different units, unnecessary repetition is avoided by keeping a complete listing of Replacement Sets in a Replacement Set Inventory. In this inventory, any particular Replacement Set can retain whatever number it was first given.

Sampling from a Replacement Set is ordinarily thought of as being done randomly, but the exact sampling procedure must be determined by the purposes of a specific field-test design. The cells in an item form function as sampling spaces which may be grouped in various ways and sampled according to various schemes, depending on what estimates one desires to make. Maximum flexibility is ensured by handling the sampling procedures separately from the item forms, in an "Assignment Plan" compiled on each occasion that a group of item forms is assembled for use in a particular test administration.

Stimulus and Response Characteristics

Sections I through IV of the item form are descriptive and analytical. These sections are important in making interpretations of the domain of items, in analyzing the functional relations between items and cells within the item form, in analyzing the functional relations between items and cells of different item forms, and in making generalizations beyond the item domain.

Of primary importance are the sections Stimulus and Response Characteristics and Cell Matrix. These are intended to describe and justify whatever behavioral analysis may underlie the properties or characteristics utilized in structuring the domain of items. Where possible these characteristics are dimensionalized and given brief titles. For example, items ask for a response to be made in a certain mode, usually either spoken or written. Thus, a possible description for this dimension is Requested Response Mode and the two values it can take are Spoken and Written.

Stimulus and Response characteristics may be usefully grouped under three general headings:

1. characteristics that are constant for (common to) all cells in the item form;
2. characteristics that distinguish among cells; and
3. characteristics that are variable within cells.

Specifying the constant characteristics is the most difficult task, especially in the early stages of subject-matter analysis. The problem is not to distinguish among items within an item form but rather to distinguish among item forms within a larger domain. The features of the item form chosen for description in this sub-section reflect the characteristics of an implied larger domain of which it is a part. The larger this domain, the more features of the item form must be identified for the purpose of distinguishing it from others. And the less clearly the domain is understood (i.e., in early stages of analysis), the harder it is to identify relevant features. Although no easy ways of clearly specifying the relationships among item forms have yet been devised, this sub-section of the item form ought to contain all relevant information possible. The development of a comprehensive classification system in "Behavioral Systematics" may provide the long-range solution to this problem.

Characteristics that distinguish among cells are usually the easiest to identify because they are the main product of the analytical dialogue. These are the major characteristics that the writer and evaluator suspect may have important effects upon performance. For the same reason, within-cell-variables are also not difficult to specify. These are the variables over which the writer and evaluator expect performance to generalize.

Cell Matrix

The Cell Matrix serves two functions: (1) it provides a summary of the information presented in greater detail under Stimulus and Response Characteristics, thereby making it easier to trace the relationships among the variables and the connection and distinctions among the cells; the two sections, therefore, complement each other; and (2) it assigns an identification number of each cell to coincide with the cell numbers used in the Replacement Scheme. The Cell Matrix is, therefore, an important link between the various sections of the item form.

Cell matrices may take different forms depending on the number and kind of dimensions used to stratify the domain. These dimensions may be completely crossed, or some may be nested within others. Sometimes blank cells will occur when combinations of item characteristics do not logically or naturally occur together.

Although the Cell Matrix has been presented here far along in this discussion of sections of item forms, it is most often one of the first sections of the item form to be written. Of course, it cannot be easily interpreted at an early stage in the production of the item form by someone other than its author. Yet it provides a convenient way of organizing the information with which he is beginning to work. Further, it sometimes happens that in laying out the Cell Matrix, item classes that have not previously been considered are discovered. These new cells are often unusual and interesting; manipulation of cell matrices thus provides a useful heuristic for extension of the domain.

Identification of the Item Form

The two beginning sections of the item form give information intended to facilitate rapid identification of its most essential features. The Item-Form Number uniquely identifies it for mechanical storage and retrieval. The Title is used for purposes of preliminary identification and classification of the behavior and ordinarily provides little specific information. The General Description gives more information than the Title but is restricted to a brief statement, usually in one or two sentences, describing more precisely the major characteristics of the item form.

Those familiar with the technical language of behavioral objectives (c.f. Mager, 1962; Popham & Baker, 1970) might find it interesting to try translating the titles and general descriptions of the item forms in Appendix 3 into that language. Some translate easily, but others do not. In general, when the behavior defined by the item form becomes complex, the language of behavioral objectives becomes unwieldy. Thus, there is no simple, direct connection between the analysis of behavior in terms of item forms and the language of behavioral objectives. The language of behavioral objectives is just one way of describing, in general terms, the classes of behavior that are operationally defined by an item form.

Scoring Specifications

The last part of the item form to be considered is the section dealing with Scoring Specifications. These describe the properties to be used to distinguish between correct and incorrect responses. However, the problem of clearly specifying the necessary properties is often not a simple one. Throughout the MINNEMAST work a concerted attempt was made to force dichotomous scoring of items. Whether or not this should be an essential feature of the domain-referenced testing model is a matter of some debate, but in practice it has two nice qualities:

1. It simplifies the conception of the basic testing model so that one may conceive it as fundamentally a problem of estimating the proportions of people who could answer an item correctly (or proportions of items a person could answer correctly) rather than as a problem of estimating mean scores on scales of correctness.

2. It has the useful characteristic of prompting writers (and other subject-matter informants) to think with unusual care about the properties that really do enter into their judgments of correctness.

A few examples may help clarify some of the issues involved. Consider an item in which the student is presented with an ordered pair of numbers and with a grid on which he is asked to plot the point corresponding to the ordered pair. Especially when the specified point does not fall at the intersection of two grid-lines, it can be expected that subjects will not plot the point precisely. The problem is how much error to allow. Determining the error tolerance must be considered an integral part of the definition of the objective. Presumably, the allowable error tolerance in graphing an ordered pair could be reduced as the testing population changed from third-grade children, to high school seniors, to research physicists—that is, as the educational objectives for each successive level of schooling specified greater precision in graphing skills.

A different kind of example, involving verbal rather than manipulative response, is an item in which the child is presented with a three-digit numeral (e.g. 237) and asked, "What is this number?" Let us imagine that a "correct" response as specified by the writer is "two hundred thirty-seven." How then should the following responses be scored: "two hundred and thirty seven"; "two thirty-seven"; "two-three-seven"? To make decisions like these, the writer may often have to go back to a more careful analysis of the exact sequence of instruction in the unit.

Sometimes it may be hard to specify an ideal target response in advance. This occasionally happens when one concocts items that place the student in a relatively unstructured situation, where a wide variety of responses may be appropriate, and where criteria for judging the goodness of the responses are complex and vague. The MINNEMAST staff found it profitable to utilize a fairly large number of such items, or item forms, and to carry out *post hoc* classifications of students' responses. Although this was time consuming, it yielded information that often proved quite useful in suggesting leads for revision of the curriculum materials. A good example of such *post hoc* classification may be seen in MINNEMAST Item Form 23.10 (see Appendix 4, Example 4). The groupings used in this classification are somewhat arbitrary and incomplete but they do impose meaning, however pragmatically, on the data. It is important to note that such classification developed in *post hoc* analysis of data from one test administration may be used as prior criteria for defining correct responses in the next. Thus the process should converge toward development of more and more precise scoring criteria.

The principal data derived from the MINNEMAST domain-referenced testing program were proportions of "correct" responses. Often however, it was also instructive to know what *kinds* of correct and incorrect responses were being made. Not only did such knowledge suggest more specific ways of revising the materials but it also occasionally suggested new objectives of the curriculum that had not previously been considered. Therefore, a simple two-way categorization of responses into correct and incorrect was often thought inadequate, and further subdivisions of those categories were made in order to take advantage of the richness of the collected data.

In general, the MINNEMAST staff found it most convenient to develop the Scoring Specifications section of the item form empirically, through analysis of responses from successive test administrations, and not to try to write air-tight specifications before gathering pilot data. Complete Scoring Specifications, except for the simplest items, were seldom written into the item form at the beginning. Strictly speaking, however, an item form must be considered incomplete until this section is written.

EXPERIMENTAL DESIGN

Choosing an Experimental Group

To assess the effects of a curriculum, it is necessary to choose an experimental group of children to whom it has been presented. Choice of an experimental group close to project headquarters was desirable for several reasons. The main considerations were convenience and management. The MINNEMAST Curriculum essentially took the form of a teacher handbook containing directives and suggestions telling how to prepare and perform sequences of instruction. As is the case with most curricula, the teacher thus mediates between curriculum writer and child. The curriculum writer cannot influence the child's behavior directly, but the curriculum may be thought of as the writer's means of guiding the teacher's behavior. Clearly, then, one line of research and evaluation ought to involve assessment of the effects of the curriculum on the teachers. Strategically, however, it makes sense to give first attention to the effects on children of the treatment "according to the book." Can children acquire the prescribed objectives when taught the curriculum as specified? Therefore, it may be necessary, at least initially, to provide guidance and consultation for the teacher beyond what is provided by the curriculum alone. The question of whether the curriculum by itself adequately guides teachers can be investigated later as a separate issue.*

---

* This argument may have a logically convincing ring, but the situation is much more complex than it sounds. The roles of teachers and developers need to be carefully re-examined: the production-dissemination model, on which MINNEMAST and most of the NSF curriculum development projects of the time were based, may be a poor model for change in education. These issues will be raised again in Chapter III.

It is probably impossible, and perhaps undesirable, for the curriculum to constrain the teacher severely. Certainly, the form of the MINNEMAST Curriculum precluded such extensive influence. Many decisions concerning classroom management, pacing of activities, and motivational strategies, for example, were left for the teacher to make when implementing the lesson plans.

In general, it is not easy to keep track of whether what a teacher is doing represents a satisfactory implementation of the curriculum. In MINNEMAST, to ensure that the curriculum was taught as intended by the writers, informal supervision was provided by a member of the writing staff (or someone of similar training and orientation) who could spot the more obvious violations of the written curriculum and provide appropriate consultation to the teacher.

The MINNEMATH Trial Centers had been set up with this problem of supervision in mind. They were to provide preservice and inservice training for teachers of the curriculum and to provide consultative help in classroom instruction. However, the extent to which the members of this loosely organized consortium succeeded in carrying out these aims was difficult to assess. On the other hand, teachers in the Twin Cities area were in continual contact with writers and implementation staff. Many of the personnel responsible for developing the units were involved in local supervision or were easily accessible for consultation. Moreover, the evaluation staff could keep itself apprised of the situation in the schools, take note of exigencies that might influence the instructional outcomes, and provide general supervision for the test administration.

Testing considerations were also important in deciding to use only Twin Cities schools for the domain-referenced evaluation. Administrative problems increased with each new center added to the experimental population. The individualized nature of the testing required the hiring and training of examiners on each location, the construction and transportation of sets of tests and test objects for each location, and the coordination of testing with differing school schedules. In addition, extra time and effort were necessary to ensure that examiners carried out the testing procedures as prescribed.

By restricting the experimental group to local schools, only one set of examiners was needed, and their training could be comprehensive and efficient. Supervision of examiners was easily accomplished by occasional observational visits and by daily scrutiny of completed tests. The construction of tests, a time-consuming task, was not further complicated by the necessity of preparing many test kits and of ensuring their arrival at many distant locations in time for administration. Finally, coordinating testing with teaching could be done without elaborate communication systems, since daily contact could be maintained with schools and testing arrangements could be quickly modified to meet local contingencies.

Generalization from this local experimental group to more general populations, of course, must be made with caution. The experimental group reflected a crude cross-section of elementary school classrooms throughout the country. Both urban and suburban schools were represented, and a wide range of socio-economic backgrounds was included.

The experimental group could have been stratified to give independent estimates of performance for various sub-categories of children. The problem is parallel to that of generalization beyond an immediate domain of test items. To the extent that the experimental group can be stratified along various dimensional characteristics, extrapolation beyond it would be possible.

But the cost of such elaborate experimental designs did not seem worthwhile at the start. Initially, the important question was whether the curriculum could be taught to any group at all. A system of successive approximation was considered the best way to proceed: improve the curriculum until it successfully achieves its objectives with a broadly representative, but conveniently located, experimental group; then begin experimenting with other groups and continue modifying the curriculum until its generalizability across the whole target population has been fairly well insured.

Each participating teacher usually taught only one class during any particular year. Because teachers' classroom strategies varied, each class received somewhat different treatment. One way of taking account of these treatment differences is to give equal weighting to each class in the test results. To this end the group was stratified by school class. With the sampling design used (described in detail later), although it would have been possible to make comparisons among schools or among classes and to further stratify the group on the basis of aptitude measures or socio-economic indices, this was not done.

Testing of the experimental group commenced soon after a teacher reported to the evaluation staff that she had completed the teaching of the relevant unit. Since each class in the experimental group did not complete the unit at exactly the same time, testing was done with fewer examiners than there were classes. Typically, three or four examiners proved sufficient for 12 to 16 classes. Where possible, testing was completed within a period of one or two weeks so that no lengthy period elapsed between the completion of the unit and the testing of any child.

No Control Groups

No control groups were utilized. When one is primarily interested in finding out what the curriculum can do and whether it satisfies its own objectives, control groups are not useful. Where the objectives of the curriculum differ markedly from the usual so that both the content and its sequencing are unique, comparisons with other curricula may only be legitimate over extended periods of instruction, such as the whole of elementary schooling. Such summative comparisons take too long and do not provide the detailed information useful in trouble-shooting the curriculum during its development. On the other hand, it can be argued that there is some worth in knowing that the curriculum really does teach something different. For example, if there is a suspicion that some objectives of the curriculum might be met were no instruction given at all, then there might be value in choosing a group of children who have not been taught that topic formally and testing their knowledge for comparison. Effects of explicit instruction in such things as "conservation of quantity" and other components of MINNEMAST lessons derived from theories of Piaget might well fall into this category.

In general, the use of control groups makes most sense when a comparison of two specific curricula can be made. It makes little sense to try to compare MINNEMAST with any and all other curricula. But there might be considerable value in exposing separate groups of children to the A.A.A.S. Curriculum *Science—A Process Approach*. Each group could then be tested on the objectives for both curricula. A comparison of performance on those objectives unique to each curriculum, and those common to both, could indicate something about the strengths and weaknesses of each curriculum. But this kind of comparison seems most profitable after initial formative evaluation of each curriculum has been made separately.[7]

Post-Testing Only

The MINNEMAST domain-referenced evaluation involved post-testing only. There is much to be said for also carrying out pre-testing using the same domain of items so that specific statements can be made concerning the increments in learning as a direct result of the curriculum. In addition, pre-testing can reveal the extent to which various forms of behavior are already in the children's repertoires before instruction begins. However, against these advantages must be set the disadvantages.

Complete pre-testing doubles the time and labor of students, testing personnel, and data handlers, and the trade-off against coverage of detailed objectives involves a serious competition. A more subtle strategy is to embed certain pre-tests in post-tests of earlier units. The MINNEMAST evaluation did not take sufficient advantage of this strategy, partly because the strategy was not explicitly recognized at an early stage and partly because of the nature of the production schedule, which made coordination across units difficult.

---

[7] A miniature example of this kind of approach, utilizing a domain-referenced testing system to compare several approaches to teaching fractions, may be found in a thesis by Sension (1971).

Ideally one might test after each unit using samples from all of the domains for the whole curriculum. In early stages of development this would be impossible because the domains for the entire curriculum would not yet be defined. It would also be time consuming and wasteful. Sometimes, we know fairly well what behaviors may or may not be in the repertoire of the experimental population, and it is foolish to test automatically unless there is good reason to think the information obtained will be worth the effort.

If it were possible to establish a hierarchy of item-form cells, then there might be an empirical basis for their inclusion in a given test administration. The assumptions on which such a hierarchy would be employed are: (1) that if behavior is displayed at some point in the hierarchy, then all behaviors lower in the hierarchy would also be displayed; and (2) that if behavior is not displayed at some point in the hierarchy, then neither will behaviors higher in the hierarchy be displayed (c.f. Gagné, 1967). The implication of this would be that item-form cells from the hierarchy ought to be excluded from the testing until there is some possibility that children can display the behavior and then be included until most children are displaying it. This evaluation strategem was not emphasized by the MINNEMAST staff.

Item-form cells may be included in the domains for a particular unit for the purpose of investigating the limits of generalizability of the behavior, even though the writers might not consider these cells as objectives of the unit. It may then be possible to compare the results obtained for different but similar item-form cells used in testing different units. For example, Unit 16—"Numbers and Measuring"—taught the use of three-place "T-notation" for representing numerals. Thus, it represented 394 as $3T \cdot T + 9T + 4$. A domain of items was designed to test knowledge of this representation scheme. Later, in Unit 24—"Change and Calculations"—index notation was introduced so that the representation became $394 = 3T^2 + 9T + 4$. A domain of items was designed to test knowledge of this new representation scheme. Neither domain was included in the testing of the other unit, but the parallelism of some parts of the two domains allowed a comparison of the response patterns. In this case, the comparison revealed that in both units many children did not cue to such symbols as $T \cdot T$ and T^2, but instead merely to the order of the digits (see Appendix 4, example 2).

TEST CONSTRUCTION

Matrix-Sampling Design

The word test has many usages, and it is necessary to clarify its meaning in domain-referenced evaluation. Any occasion on which a set of item forms was used to generate items that were administered to a population of children was called a testing or test administration. The ordered set of items assigned to a particular child was called a test. But since each child usually received a different set of items, there were as many tests as children sampled. The individual tests were of little interest in themselves; they were simply the outcome of a sampling design aimed at producing estimates of the performances of the population of children over various domains of items. Usually, estimates of performance were obtained for each item-form cell in the domain.

The simplest design, if it were possible, would be to assign *all* items in a cell to *all* children in the population. The data obtained from administering all of M items to each of N children can be cast into a person-by-item matrix of the form shown in Figure 3.

Persons	Items				
	1	2	3	4 . . . M	Row Totals
1					
2					
3					
4					
. . . N					
Column Totals					

Figure 3: Person-by-Item Matrix.

Let the entries in this matrix be either "ones" (for correct responses) or "zeros" (for incorrect responses). Summing for each row or column gives the number of correct responses in that row or column. These totals can be converted into proportions by dividing by the number of entries in the row or column. If person effects are of interest, then the row proportions can be examined; if items effects are of interest, then the column proportions can be examined; if one is interested in an overall effect—a general measure of how well everyone did on all the items—one can add up all the correct responses in the entire matrix and form a proportion by dividing by the total number of entries. The MINNEMAST research and evaluation staff focused on this as the measure of greatest practical interest for purposes of initial, formative evaluation. Roughly speaking, the larger this proportion, the more effective the curriculum may be said to have been in teaching the knowledge or skill represented by the set of items in the item-form cell to the population of people represented by the set of persons in the group who used the curriculum. Let us denote this general measure P_{cg}, the proportion of correct responses (P) in the matrix formed by the set of items in the item-form cell (c) and the set of persons in the group (g). Since both c and g are finite, it would be possible to actually obtain the measure P_{cg} by administering all the items in c to all the people in g, but for virtually all practical purposes, P_{cg} is a parameter to be estimated. The development of efficient sampling designs for estimating P_{cg} for a large number of cells simultaneously (and for obtaining estimates of person and item variance within the matrices at the same time) is a topic of some complexity. (See Husek & Sirotnik, 1968, for a useful introduction to the literature.)

The procedure utilized in the MINNEMAST Project was to randomly sample a specified number of items from the item-form cell, randomly sample the same number of children from the group, and then assign each of these children one of the items. Items were sampled with replacement and children sampled without replacement. That is, each row could have only one entry chosen from it, but columns could have more than one.

There are three somewhat simpler procedures for sampling a cell-group matrix than that utilized in the MINNEMAST testing.

1. One could sample entries without replacement. In practice this could be done by drawing an item at random, drawing a person at random, and assigning them to one another. Both the person and item would then be replaced before the next draw, but the same item could not be assigned to the same person more than once.
2. One could sample a unique set of intersections of rows and columns. That is, both items and persons would be sampled without replacement, thus eliminating the row occupied by the person and the column occupied by the item from further sampling.
3. One could sample according to various block designs in which rows and columns were allocated specified numbers of replications.

Block designs represent a level of sophistication beyond that required in an initial formative evaluation, but either of the first two sampling plans would have served well. From the present vantage point, it is hard to reconstruct the arguments in favor of the hybrid sampling scheme utilized in the MINNEMAST testing, but the differences among the distributions of the estimates obtained from these alternative methods would probably be very small.

After a sample of items and people has been drawn for the purpose of estimating a particular item-form cell proportion (P_{c1g}), the people are replaced, and the process is repeated *independently* for the purpose of estimating the next (P_{c2g}). The independence of these estimates is very important, because that is what permits descriptive flexibility, enabling one to compare proportions or distributions of proportions among any chosen item-form cells, and to combine proportions to create summaries of overall performance across several cells of item forms.

Assignment Sets

In MINNEMAST, the item-form cell was ordinarily used as the basic sampling space for items. However, sometimes one may wish to combine items from several cells into a larger set. This gives rise to the notion of an Assignment Set.

There are two main reasons for creating special-purpose assignment sets:

1. One may wish to obtain general estimates of performance over several related cells without spending the time and effort required to draw enough items from each cell to obtain a stable estimate of P_{cg}. For this purpose, the cells may be grouped together in an assignment set for which a special stratified sampling plan is specified. For example, two items might be drawn at random without replacement from each cell and the sets of items obtained in this manner might be assigned to a sample of students in the same way as if they constituted a sample of items from a single cell. This would yield a general estimate of performance over the domain formed by the collection of cells, in which the behavior represented by each cell would receive equal weight.
2. One may wish to avoid assigning two or more very similar items to the same student. Sometimes items from related cells may be so similar that cuing effects may be predicted. If one wishes to prevent such effects from occurring, the cells in question may be grouped into a single assignment set for which students are sampled without replacement.

The Assignment Plan

Sampling directions for a particular testing are drawn up in an Assignment Plan. One part of the Assignment Plan specifies the relevant experimental group of students and how it is to be stratified, if at all, for sampling purposes. A second part specifies the item-form cells to be used, the ways of composing special assignment sets (if any), the numbers of items to be drawn from each cell, and the maximum number of items allowed each child. Finally, the Assignment Plan must specify the exact procedure for assigning the items to the people.

In the MINNEMAST evaluation the group of students was stratified by school-class. Classes were identified by school name, class number, and teacher's name, and for each testing each teacher submitted a list of children who had been presented with the relevant unit. This sufficed for the purposes of MINNEMAST, but for long-term studies students should be assigned unique serial numbers to make it easier to select groups with more extended histories, perhaps across several years and in several different classes.

The procedure for assigning items to students was as follows: The sample of items was first generated. Letting i be the number of items and c be the number of classes, if i happened to be an even multiple of c, then $\frac{i}{c}$ resulted in a whole number x which indicated the number of children to be sampled randomly from each class and each randomly assigned to an item. If i was not an even multiple of c, then $\frac{i}{c}$ resulted in a whole number x plus a remainder y. This meant that after sampling x children randomly from each class there were still y more children to be chosen. These were obtained by selecting one child from each of y classes chosen randomly from the c available classes.

Sample Size and Reliability

How many items need to be drawn from a given item-form cell in order to obtain a reliable estimate of P_{cg}? The details of this problem have not yet been thoroughly worked out, but its general outlines are fairly clear. The cell-group matrices are finite, but usually they are rather large. In placing bounds on predicted sampling variability, it is probably safe to think of the worst possible case as being an infinite matrix in which the value of each entry is determined independently of the others. Of course, the entries are not independent; in fact, items are classed together in item-form cells precisely because one expects the behavior of students to generalize across them. Similarly, students are grouped together in strata because one expects the behavior of students to possess certain commonalities. The effects or homogeneity of items, and of students, must be to decrease sampling variability in comparison to the hypothetical worst case.

A rough idea of the effects of sample size on the reliability of the proportions in the "worst case" may be formed by considering the confidence limits shown in Table 1. As a practical operating rule, sample sizes of approximately 30 seem to be fairly satisfactory. In the MINNEMAST evaluation, this was the rule of thumb followed.

Table 1: Confidence Limits (95%) for varying sample sizes with $p = 0.60$, assuming normal distribution of the proportion.*

N	Lower Limit	Upper Limit
10	.25	.95
15	.32	.88
20	.36	.84
25	.40	.80
30	.43	.77
40	.45	.75
50	.49	.71
100	.52	.68

---

* The formula used for calculating these values is given in Walker and Lev (1958), p. 249.

In practice, the question of the reliability of the estimated proportions is largely an empirical matter. The main issue is whether or not the obtained proportions fall into orderly and interpretable patterns, and in general they did. Some representative data are presented in Appendices 4 and 5.

Limiting Number of Items per Child

As a consequence of the matrix-sampling design, each child could be assigned a different number of items. Some might receive none at all; others might, by chance, receive a large number. To prevent individuals from receiving an unmanageably large number of items, it may be necessary to set an upper limit to the number of items which can be assigned to any one child. Provided that the proportion of children who reach this limit is low, the independence of the estimates should not be greatly affected.

For MINNEMAST testings, 10 items per child was usually the specified maximum. But the value chosen depends on several considerations. Older children might be allowed more items than younger children. The number of manageable items also depends a good deal on the nature of the tasks. Some of the MINNEMAST domains involved very complicated, time-consuming activities. Others were simple and involved less time and effort.

In the MINNEMAST procedure, if a child reached his limit of items before sampling for all assignment sets was completed, he was eliminated from sampling for the remainder of the assignment sets. The result of this was that the group of children from which samples were drawn early in the sampling process was not quite the same as the group of children from which samples were drawn later in the sampling process. If a large proportion of children were removed from the population through this process, serious problems of inference would arise. The MINNEMAST testings were moderately successful in keeping the proportion of children reaching the set limit low. Table 1 of Appendix 4 illustrates a case where an upper limit of 5 items per child was imposed for each separate part of a test administration.

In general, careful consideration must be given to the interplay between the number of cells in the domain, the number of items generated per cell, the size of the experimental group, and upper limit of the number of items assigned each child. Decisions cannot be made about one without affecting the others. If the size of the experimental group is fixed in advance, then the others must be adjusted to accommodate. If the experimental group is small and the domain large, then it may be necessary to restrict testing on any given occasion to only certain cells from the domain.

Randomizing Order Effects

Systematic carry-over effects from one item to the next may be produced if items are arranged in the same sequence in every child's test. Such effects are usually undesirable, because we wish to study items independently of each other. They may be avoided by randomizing the order of presentation of items to each child. This was the practice adopted for MINNEMAST testings.

There may be times when the effects of particular orderings are of interest. These may be studied by arranging specific orders of items as explicit variables by means of a special assignment set. No attempt was made to study such orders in the MINNEMAST testings.

Practical Procedures in Test Construction

Test construction in accordance with a matrix-sampling design tends to become time consuming and tedious unless considerable effort is devoted to designing efficient procedures. In theory, domain-referenced testing is well suited to computerized generation of the printed materials and matching of items and children.

In the early days of MINNEMAST testing, it was not possible to anticipate all of the variations in sampling schemes and test-item formats that were to come. As a result, the first computer programs were not applicable to later domains without considerable modification. It quickly became apparent that changes in the programs could not be made fast enough to provide complete print outs in time to meet testing schedules. Consequently, complete reliance on the computer was abandoned and more flexible techniques substituted.

One aspect of test assembly which remained essentially unchanged throughout the project was the manner in which children were sampled and assigned to test items. Therefore, it was feasible to use a computer program for this specialized purpose. Given an assignment plan, the computer printed out a set of labels each showing a child's name, the identification number of the item assigned to him, and other identifying information such as his class, teachers, and school. All of the labels for items assigned to a particular child were printed out together and in the (randomized) order in which the items were to be administered. The names of the children were printed out alphabetically within classes.

The items themselves were assembled by hand rather than printed out by the computer. A ditto master was made for each item-form shell and the necessary number of copies produced. Individual items were then constructed by filling in the blank spaces in the item-form shells according to the replacement scheme of the item form. Sampling from each replacement set was done without the use of a computer. Instead, either a table of random numbers was consulted or numbered counters were drawn, lottery style, from a cup. Typically, the size of the replacement sets was small enough to make these procedures efficient.

Items generated from an item-form cell were referred to as replications from that cell and labelled with consecutive numbers, referred to as replication numbers. Therefore, each item was uniquely identified by a sequence of numbers specifying the item form, cell, and replication. (In practice, replication numbers were assigned successively during replacement of the first variable in the item-form shell and preserved for all other replacements.) The numbers thus given to the items themselves could then be matched to the item numbers on the computer print out labels.

When all of the items specified by the assignment plan had been generated, the computer print out labels were attached to them and they were collated into tests for individual children. Working from the labels, each indicated item was located, and its associated label was attached. By following the order in which the labels were printed out, the items were systematically collated into tests for each child, with items within each test arranged according to the assigned order and children within each class arranged alphabetically. Finally, the tests were put together in bundles for each class within each school.

All the necessary test objects must be constructed, too. It is worthwhile starting to assemble these sometime before constructing the items and, in fact, many were put together at the time when item domains were being constructed. By making the objects as early as possible, difficulties in construction or supply could be overcome without delaying test administration. Test objects not attached to individual items were assembled into a Test Object Kit for each examiner. Reusable materials were always put into this kit rather than attached to the items.

The assembly of the tests and the test object kits was a task assigned to several part-time undergraduate assistants. Sometimes, the construction of test objects needed the personal attention of a member of the evaluation staff, but for the most part, the procedures were routine enough to require little outside help or supervision.

In general, the foregoing procedures make up a workable production system that does not depend heavily upon sophisticated hardware. To go beyond these procedures will require the development of very general and flexible computer languages for manipulating verbal and symbolic materials in item forms.[9]

Another function in which computer support would be a great help in the long run is that of indexing, cross-referencing, and storing item forms and performance data. At the time the MINNEMAST Project terminated, the research and evaluation staff had just begun to grapple with this. The problems become formidable when one begins to contemplate the long-term development of item-form banks.[*]

---

[9] One such language is currently under development at the University of Minnesota under the direction of Rod Rosse, Department of Psychological Foundations of Education.

TEST ADMINISTRATION

Test items for the evaluation of MINNEMAST Units were designed to be administered individually. Individualized administration of items was not a necessary component of the general evaluation model and, had the project proceeded with the development of units for the upper grades, individual testing might have been to some extent supplanted by group testing.

Individualized testing necessitated the hiring of several people to administer the tests. It was important that the examiners appreciate the necessity of following the item directions as given and not interpolate procedures and comments of their own. Therefore, where possible, people with some training in testing were hired as examiners. Usually, they were undergraduate students in psychology or education.

It was possible to arrange for a fairly continuous schedule of testing during the school year. Within any one grade level, different classes completed a given unit at different times, often a month or more apart. Similarly, from grade to grade, units were completed at different times throughout the year. Consequently, examiners could be kept busy throughout the year testing units at various grade levels.

Because of the staggering of testing, a large number of examiners was not needed. Three or four examiners were able to carry out the testing of approximately eight units each year with about 400 children involved in each testing. Training sessions for examiners were held by members of the evaluation staff to explain the evaluation model and to describe the general procedures to be followed.

Examiners were advised to spend some time in small talk with the child before beginning testing, and each test included a sample warm-up item to ease the child into the test.

The examiner was instructed to sit on the same side of the table as the child. This served two purposes. Seated next to the child, the examiner was in a better position to observe his activities. Further, in this position the child was less able to observe the facial expressions of the examiner and therefore was less likely to obtain inadvertent cues about his progress.

Examiners were instructed not to depart from the exact words of the items once test administration had begun. If for some reason they found it necessary to add some comment or question of their own, they were to record their words on the item so that due allowance could be made for them later. Questions from the child about the item were answered by saying simply, "I'll read it again," and then repeating the question. The same procedure was adopted if the child failed to respond within 30 seconds. If the child still failed to respond after another 30 seconds, or when it appeared that he had completed his response, the examiner asked, "Shall we go on to the next one?"

---

[*] An excellent operating example of a computer-based item selection, generation, and printing system is that produced at the Center for the Study of Evaluation, the *System for Objectives Based Evaluation—Reading* (Skager, 1971).

Recording of responses was done in as detailed a manner as possible. Spoken responses were recorded verbatim. Manipulative responses were recorded by a detailed written description and the use of diagrams. Written responses were made on sheets of paper which could then be stapled to the relevant item. The important thing was to ensure that an outsider could accurately reconstruct the essential features of the response. The examiner made no judgments about the correctness of the response.

Sometimes it happened that a child to whom test items were assigned was no longer attending the school or was absent during the whole of the testing. When this occurred, his test was reassigned to another member of the same class. In reassigning a test, a random selection was made from those children who had not been assigned any items. When all children had been assigned tests, selection was made randomly from those with the fewest items assigned. Since one or two such absences per class were not infrequent, the reassignment of tests was important in preserving the sample size.

Training of examiners also included detailed mock administration of the items. It is important for the examiner to practice with the items and test objects before testing begins in order to reduce the chances of inadvertent mistakes in administration. Each day, examiners brought their completed tests back to the evaluation staff for "de-briefing." Quick action could then be taken if it turned out that some items were not being administered correctly. Wrongly administered items were excluded from the compilations of results.

Another form of supervision of examiners was occasional visits to the sites of testing by a member of the evaluation staff. Spot checks on examiners early in the testing schedule enabled incorrect procedures, such as rephrasing of scripts and incomplete recording of responses, to be remedied. By taking trouble early to ensure that an examiner followed the rules, testing was able to proceed smoothly with a minimum of operational problems.

On a few occasions, recording methods other than written descriptions (sound or video tape) were used. In general, however, these created more problems of data processing than they solved in data collection.

INTERPRETING RESULTS OF TESTING

Categorizing Item Responses

Testing produced an enormous quantity of raw data. For a single MINNEMAST Unit it was usual to administer between 1,000 and 3,000 items. The item responses had to be scored and codified in ways that preserved the interesting and important facets of the responses and allowed those facets to be easily seen and understood.

The responses were almost never of a kind which would lend itself to computer scoring. Instead, scoring was done by hand by members of the evaluation staff. Since scoring criteria were often not established in advance of testing, scoring usually required the establishment of categories of responses for each item-form cell. When possible, a correct-incorrect dichotomy was established. But often this was insufficient, and further subcategories were developed on the basis of salient differences among the responses.

No set procedures were established for presenting data. Each staff member drew up tables and made statements about them in whatever way he considered most amenable to the data. Toward the end of the project, it became usual to present the data both in tables showing frequencies of various categories of responses and also in complete item-by-item listings of the actual responses (condensed where necessary to enable them to be reported in this way) arranged according to the categories shown in the table. This arrangement allowed others to check the accuracy and suitability of the categorizations (see Appendix 4, Examples 3 and 4).

Statistical Inference

The basic statistic utilized in presenting the test results was a proportion reporting performance on the sample of items from the item-form cell by the sample of children from the experimental group. This proportion was treated as an estimate of the corresponding parameter of the cell-group matrix, viz., the expected performance on all items from that item-form cell by all children in the experimental group. Responses were scored dichotomously so that the proportion represented correct or satisfactory responses. Related analyses produced proportions for subcategories of responses, and these were often reported along with the proportion of correct or satisfactory responses. Proportions for these categories of responses were treated in the same way as proportions of correct responses—as estimates of the corresponding parameters.

The proportions of correct responses for each item-form cell were usually cast into a table following the characteristics of the cell matrix of the item form. This table was then examined for patterns that might be attributed to the dimensions of the matrix. No inferential statistical tests were applied to these tables; inductions were simply made from the data by searching for consistent patterns. For example, in the data from Appendix 4, Example 1, which appear in Table 2, the consistent similarity of the proportions across columns shows large effects on performance related to ranges of numbers and the conjunctions of order relations, but no effect related to response mode.

Table 2: Proportions of Correct Responses for the Cells of Item Forms 9.7 and 9.8: Producing Number (×) Satisfying Given Relations to Specified Numbers (b_1 and b_2)

		Response Requirement				
Item Form	Response Form	(1) $x > b_1$	(2) $x < b_1$	(3) $b_1 < x < b_2$ $b_2 > b_1 + 1$	(4) $b_1 < x < b_2$ $b_2 = b_1 + 1$	(5) $b_1 < x < b_2$ $b_2 < b_1$
9.7	Spoken	$\frac{18}{20}$	$\frac{9}{10}$	$\frac{16}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
9.8	Written	$\frac{18}{20}$	$\frac{10}{10}$	$\frac{11}{20}$	$\frac{0}{20}$	$\frac{0}{20}$
	Totals:	$\frac{36}{40}$ (.90)	$\frac{19}{20}$ (.95)	$\frac{27}{40}$ (.68)	$\frac{1}{40}$ (.03)	$\frac{2}{40}$ (.05)

This emphasis on analysis in terms of patterns of responses to sets of items with related stimulus and response characteristics is the most powerful product of domain-referenced evaluation. Where an objective is represented by a single item, and students have difficulty answering it correctly, it is easy to seek explanations for their performance in incidental or artifactual characteristics of the item. But if the objective is defined by sets of items with related stimulus and response characteristics, the functionally effective variables emerge clearly through analysis of the overall pattern of difficulties. Examples of such pattern analysis are illustrated in Appendices 4 and 5.

Standards of Performance

Desired standards of performance against which actual performance could be compared were not established in advance of testing. In general, there was little experimental evidence or theoretical rationale to guide the setting of desired levels of achievement, and staff members often disagreed, some being content that a few children attain some objectives and others being dissatisfied if most children did not attain nearly all objectives. Consequently, proportions were reported and patterns in the cell matrices examined, but without the application of any set standards of performance. Instead, the writers were left with the problem of deciding *ex post facto* whether they were satisfied with the results. A more structured rationale for interpreting results, agreed upon by all project members, and perhaps periodically debated and revised, would have been useful.[10]

Interpreting the Results and Revising the Curriculum

Evaluative reports typically consisted of three sections: at the lowest level of abstraction was a section showing relative frequencies of specific responses grouped into analytical categories within each separate item-form cell (see Appendix 4, Example 2); at the next level, proportions were reported in such a way as to allow several related item-form cells to be examined at the same time (see Appendix 4, Examples 1 and 3); and at the highest level, a short interpretive summary was prepared, pointing out the main characteristics of the data (see Appendix 5).

Once the evaluative report had been prepared, discussions were held with the writers concerning its significance. Typically, a general staff conference was scheduled to examine and discuss the report. Appendix 5 illustrates the typical quality and complexity of the data presentation. It was at the point of revision that the domain-referenced evaluation activity ran into most severe difficulties. In general, the data presented in the evaluation reports and discussed in staff meetings did not stimulate and guide revisions of the units effectively. A body of theory and a set of procedures to guide revision did not accumulate. Some very general organizational problems of the MINNEMAST Project made it difficult to utilize the data—problems representative of those encountered in many other large-scale curriculum development projects. Recommended organizational strategies for future projects, designed to circumvent some of these problems, are presented in the next chapter.

A NOTE ON THE COST OF THE ACTIVITY

The cost-benefit of domain-referenced evaluation is of considerable interest, because even under optimum project organization, the critics contend that procedures may be so expensive as to be unworkable. During its peak years, 1967–8 and 1968–9, the MINNEMAST domain-referenced testing activity involved about one-quarter of the professional staff. This proportion was estimated by ignoring expenditures for administrative-clerical staff, artistic-reproduction staff, and trial-center personnel (all of whom may be thought of as forming a central core, supporting production and evaluation activities alike) and comparing the salaries of writers, editors, and their assistants to the salaries of psychologists and their assistants. During fiscal 1968, for example, salaries for the writing staff totaled approximately $188,500 and for the evaluation staff approximately $54,000. During that year, eight units were evaluated at a cost of roughly $7,000 per unit.

---

[10] This problem is similar to that of interpreting the results of the National Assessment of Educational Progress (see Stake, 1970). In general there is an interesting parallel between the goals and strategies of the MINNEMAST domain-referenced evaluation and the goals and strategies of the National Assessment (c.f. Merwin & Womer, 1969).

CONTRIBUTIONS OF THE MINNEMAST PROJECT

In many ways the MINNEMAST project was ahead of its time. It pioneered basic notions about content, structure, and process in elementary mathematics and science that have since been incorporated in curricula throughout the country. It opened the way for strategies of interdisciplinary development, dissemination, and evaluation, and produced an important generation of professional educational developers. Much of this intellectual endeavor may outlast the elementary curriculum materials themselves that formed the concrete product of the work. One purpose of this monograph has been to document some of the results of what was a most exciting, cooperative venture.

Chapter III

CONSIDERATIONS IN THE DESIGN OF FUTURE SYSTEMS OF CURRICULUM CONSTRUCTION AND EVALUATION

FACTORS THAT INFLUENCE THE EFFECTIVENESS OF CURRICULUM DEVELOPMENT PROJECTS

In the preceding chapter we noted that some of the organizational characteristics of the MINNEMAST Project constrained the effectiveness of its curriculum development and evaluation activities. These characteristics represent potential problems encountered in virtually all major curriculum development projects. This chapter discusses some of these potential problems, not only as they relate to the MINNEMAST Project, but also in terms of how they may limit the effectiveness of curriculum development and evaluation in general. In addition, recommendations are made concerning how these problems might be overcome, with direct application to domain-referenced evaluation.

Need for Clear, Cumulative Overall Strategy

Major problems often arise because of the absence of clearly specified and well-functioning organizational mechanisms for planning and supervising the overall activities of a project. Ideally, the general strategy should be cumulative in the sense of making continual progress toward the achievement of defined goals, each successive step building on previous accomplishments. In any large project this ideal is difficult to achieve, because experience constantly suggests changes in both goals and procedures.

A helpful step toward developing a cumulative strategy, however, is to preserve good records of issues met and decisions made. A project archivist should be responsible for maintaining files on issues, proposals, decisions, implementations, revisions, and rationales, cross-indexed and summarized so that when new issues arise they can be treated in the context of relevant history. These files should contain everything of relevance to the developmental effort, including, where possible, staff memoranda and occasional papers, minutes of meetings, details of decisions taken and reasons for them, designs of instructional sequences and lessons, and reports on evaluations. These archives may serve as an adjunct to the library and enhance its general educational function.

Potential ambiguity can be removed by maintaining a public blueprint of a project describing its philosophical rationale, the general operating procedures of the organization, and the general characteristics of the intended products. As the project proceeds, this document can be revised and extended but always kept in the forefront of planning and review. Revision of the document is occasioned by deliberate decisions that the goals of the project should be altered. Extension of the document involves interpretive explications according to deliberate decision about meaning and intention. In a sense, the whole effort of the project is directed at operationalizing the blueprint.

Turnover of Staff

Cumulation of effort may be hampered by turnover of staff, traceable to three main frustrations: (1) conflict among staff members arising from vaguely defined goals and inadequate communication and decision processes; (2) relative instability of the job; and (3) inadequacy of the available professional and personal rewards.

The first of these problems is often connected with the lack of clear overall organizational strategies. People who do not think the organization is doing what it should be doing, and who do not believe that their views have received reasonable consideration, will either attempt to seize control or, failing that, will leave.

Appointments of committed staff may be unstable because of uncertainty that funds will be available for each coming year. Faced with this uncertainty, many staff members who might be disposed to spend a major proportion of effort on a project might seek advancement in their careers and stability in their employment elsewhere. One consequence of this can be that many of the key staff work only part-time on a project, which can compound organizational problems in communication and coordination. A central group of full-time staff invested with responsibilities for leadership and coordination in a project can produce a much more effective and efficient operation. A long-term funding arrangement allows at least a few people to commit themselves to such a role for an extended period.

Curriculum projects frequently enlist the help of numerous scientists and mathematicians. These people may initially volunteer their time and effort as occasional consultants, summer writing-session participants, part-time staff members, etc. But scientists and mathematicians who contribute this kind of time and effort run the danger of losing their professional status and orientation, since the more time they spend away from their own direct, professional pursuits, the less likely they are to stay abreast of their field. The point is not just that subject-matter experts who throw themselves into the effort of curriculum reform may lose out professionally, but that many of them fail to make themselves available at all for fear of falling behind. There are two basic issues: (1) curriculum writers are usually not given the kinds of rewards considered satisfactory by the professional scientist or mathematician; that is, they are rewarded not for behaving as scientists and mathematicians, but as somebody else whose activities are essentially unscientific and non-empirical; and (2) being a curriculum writer in any case involves learning new skills which are not those of the research scientist and mathematician.

One suggested way of overcoming this problem is to involve not the advanced level expert, but his students. At first sight this seems an attractive proposition. Graduate students should understand the basics of their field as well as their professors, may perhaps be more flexible than their professors concerning problems of instructional design and, moreover, cost less to employ. But the same serious flaws remain. Graduate students have presumably embarked on a career in their chosen field, just as their professors. For a graduate student to take time out to study the issues of curriculum design can be threatening to his career, also. The same pressures are likely to build up for him to change his career plans or to be content with second-rate professional status.

For a curriculum development project, an alternative approach to this problem might be to use subject-matter experts not as writers but as informants; people whose behavior is *studied* by educators and educational psychologists. Detailed and systematic study of the behavior of scientists could yield the information necessary to set up domains of behavior to represent the objectives of instruction and to suggest effective teaching techniques. By sampling from pools of available experts, the demands on any one of them could be kept small.

Drag Exerted by Production on Evaluation and Revision

There are other problems of organizational strategy which may seriously affect the evaluative effort. For example, evaluation may be conceived on a grand scale, as was the case in the MINNEMAST Project, and encompass large numbers of students in classrooms all over the country. In a sense, then, implementation of a curriculum in these classrooms has less to do with evaluation than dissemination. And in that sense, the curriculum may be largely prejudged to be worthy of implementation on a large scale. Evaluation under these circumstances is bound to be relegated to a summative rather than a formative role in an organization.

This conception of evaluation is not only peripheral to the development of the curriculum, but is also a relatively static activity. That is, it may encompass such long time periods and involve materials upon which such large expenditures of time, effort, and money have been made, that there is relatively little opportunity for it to influence the production process. At most, this evaluation only serves judgment on what has been done. It does not have the characteristics of a "servomechanism." For that, a smaller, more easily controlled, more frequently applied evaluative effort is needed, and it needs to be linked closely with the developmental effort itself. In short, the evaluation of the curriculum should be central and dynamic rather than peripheral and static. This may seem a bit like putting the cart before the horse, forcing development to fit an evaluation model. But in another sense, what is being argued is the need for an integrated developmental and evaluative effort, since evaluation ought to be the means for monitoring the developmental effort and adjusting it continually to keep the project on course toward its goal.

Production schedules present another strategic problem. A project may encounter severe problems in meeting deadlines imposed by the effort to keep schools supplied with a new curriculum. As part of the production schedule, considerable effort must be spent in preparing, shipping, and handling materials. As a consequence of the constraints set by the production schedules, rapid trial, feedback, and change may be difficult to bring about. For example, the total turnaround time of development, trial, evaluation, and revision of a unit for the MINNEMAST Project was about two years. The drag that dissemination often exerts on the production process, however, may be alleviated if the two tasks are separated.

An alternative strategy might be to concentrate the efforts of a project on a small part of the curriculum and carry on the rest of the instruction by "conventional" means and thus remove the pressure of time schedules that require a specified output within a specific time period. However, this requires a longer time perspective on production output, an insistence on proceeding experimentally and incrementally, and a suspension of judgment about the curriculum until pertinent evidence is available. The effect of working on just one part of the curriculum at a time is to concentrate the effort intensively on the solution of basic issues—of objectives, of teaching strategy, and of evaluation—so that they can be tackled systematically. In this alternative approach, large-scale dissemination is viewed as a process demanding attention only after a sizable piece of the final product has been developed to a point of satisfactory application.

Empirical vs. Artistic Approaches to Teaching

Underlying the strategic problems of development and evaluation is, of course, a philosophical problem. Curriculum writers often espouse an experimental approach to teaching in which the objectives of teaching are expressed in terms such as "exposure" and "give a feel for." Emphasis may be placed on what the teacher is to perform and not on what the child is to learn. Writers may balk at requests to think carefully and systematically about learning outcomes of the instruction. Yet, such requests may frequently be made by evaluators in order to arrive at domains of behavior representing the objectives of the curriculum. Whereas writers insist on the primacy of the instruction (what you arrange for the child to do), the evaluators insist on the primacy of the objectives (what you hope the child will know as a result of the instruction).

Essentially, this conflict is one between an empirical and an artistic approach to teaching. The empirical approach emphasizes the importance of documenting instructional issues and examining them empirically. Such an approach leads logically both to domain-referenced testing and to organizational strategies which emphasize experimentation and revision. The artistic approach de-emphasizes testing in any form and emphasizes strategies for production and dissemination of materials that are judged to be important or exciting on subjective or logical grounds.

Those who take the empirical approach tend to argue that curriculum development is an open-ended process of continual trial and improvement, and emphasize formative evaluation. Those who take the artistic approach tend to think of it in terms of writing projects with clear beginnings and ends and think of evaluation mainly in the summative sense. A practical educational venture may be organized along exclusively artistic or exclusively empirical lines, or it may strive to maintain a balance between the two approaches. It is of the greatest importance that funding agencies and development organizations approach the problem of achieving a satisfactory balance directly and self-consciously in the future.

Loose Connection Between Written Curriculum and Classroom Activities

A potential difficulty may arise if the connection between the written curriculum and the classroom activities of the teacher is a weak one. From the empirical viewpoint an ideal curriculum should give unequivocal directions as to how to make certain things happen in a classroom, given that those things are known from experimentation to lead to specified learning outcomes. The written curriculum thus turns out to be a report of procedures that have actually been tried out. This implies close initial cooperation between writers and teachers in a laboratory-classroom setting. Separation of the two may create a barrier to initial tryout and evaluation of content and method. Writers may not hold themselves closely accountable for the learning of any specific group of children and teachers may not hold themselves accountable for the implementation of the curriculum according to the specifications of the writers.

AN ALTERNATIVE STRATEGY OF CURRICULUM DEVELOPMENT

Much of the discussion so far has focused on some of the problems that may prevent the techniques of the domain-referenced evaluation from contributing significantly to the development of curriculum. Some suggestions have been made for modification that would permit these techniques to contribute to more efficient formative evaluation. In the final analysis, however, what may be needed is not an overhaul but an entirely different strategy for curriculum development. The discussion that follows proposes some general characteristics of an alternative strategy based on a thorough-going empirical approach. The advantage of a basically empirical approach is that it can be opened up to leave room for artistic production but, as we have seen, an approach primarily guided by large-scale artistic production is very hard to modify so as to incorporate systematic utilization of empirical data.

The proposed strategy is applicable to curriculum development in any area (or combination of areas) in the sciences, arts, or humanities. Following the case-history theme of this monograph, however, the reference will continue to be to curriculum development in mathematics and science.

Focus on Specific School Systems

Projects should be administratively based in school systems, not in universities or research-and-development agencies that then establish transitory relationships with a variety of schools. Two important consequences follow from anchoring curriculum development in a specific group of schools. First, a real community of consumers is established, with development of the curriculum implying its actual, long-range implementation by that community. The developer is not permitted to avoid specific problems by appealing to some more abstract target population located elsewhere. Second, a direct connection may be established between curriculum development and teaching, allowing curriculum developers and teachers to closely coordinate their activities and to make it possible for both to be held accountable for the results.

Direction by R and D Specialists, Not Subject-Matter Experts

Some of the difficulties stemming from direction by subject-matter experts have already been discussed. Taking on such activities threatens the careers of those who do so by cutting the time they devote to their normal professional activities. Furthermore, subject-matter experts typically experience considerable difficulty in approaching the design of curriculum within an experimental, scientific framework. A better organizational scheme would allow the experts to continue their professional activities while serving as advisers and informants to projects directed by social scientists and educators.

Provision for Systematic Definition and Refinement of Domain-Referenced Objectives

The task of generating domains should be performed by teams of psychologists trained in procedures for extracting them from the repertoires of experts. There are three distinct components to this task. First, there should be a systematic method for locating and sampling a broad range of experts (mathematicians, physicists, biologists, etc.) to act as informants whose behavior may be analyzed. Second, there should be a method for having these same experts criticize and guide the selection of portions of the resulting domains as goals for the curriculum. Third, there should be a method for having them suggest and criticize alternative possible teaching sequences and procedures and observe them in operation.

Another group of experts, also, ought to make a direct contribution to the selection of goals. They are the sociologists, social psychologists, economists, philosophers, and other futurists concerned with educational policy. These are important resource personnel for discussions concerning the anticipated consequences of selecting certain kinds of goals and of emphasizing one set rather than another. The expertise of these persons could be drawn on in the same way as for the mathematicians and scientists—by sampling systematically from the available pool of talent in a kind of jury duty to education.

Finally, members of the community served by the school (students, parents, and representatives of community interest groups) should be engaged in the selection of goals and the placing of priorities. Procedures for achieving this in a practical and effective way are one of the toughest current problems in American educational policy. School-based curriculum development might be set up to approach this communication problem directly and intensively.[11]

The specification of objectives in the form of behavioral domains is not, of course, a once-and-for-all process. Objectives continually evolve in the light of new information about the society, the subject matter, and the process of human learning. This would require continued, systematic sampling from the pool of social planners and community representatives for regular feedback about general objectives, systematic sampling from the pool of subject-matter experts for feedback about specific objectives, and continued assessment of the validity of the domains through experimentation and naturalistic observation of students.

---

[11] Techniques for engaging the community in selecting goals and establishing priorities have been developed and field tested by the Center for the Study of Evaluations. See Klein, S. P. Burry, J., Churchman, D., & Nadeau, M., 1971; Klein, S. P., Burry, J., & Churchman, D., 1972; Hoepfner, R., Bradley, P. A., Klein, S. P., & Alkin, M. C., 1973.

Local Assessment to Provide Performance Baselines

Development should proceed in the context of a baseline inventory of the performance of current students. The purpose of this inventory is to find out what portions of the specified behaviorial domains are already being established in the repertoires of the children in the experimental group by current experience and instruction. (This may be viewed as a kind of local counterpart to a national assessment.)

With a description of what the children are currently learning and a description of what it is considered desirable for them to learn (both stated in terms of behavioral repertoires), priorities may be established for altering parts of the instructional program.

Experimental Design of Instruction

The design of teaching strategies should proceed empirically. Teaching strategies ought to be treated as hypotheses whose acceptance or rejection is to be made on the basis of empirical investigation. Consequently, the roles of teacher and curriculum developer should be tightly linked together. If the roles of developer and teacher are separated, the developer may have difficulty persuading the teacher to follow his instructions or, worse still, may become so remote from the classroom situation that his instructions are unnecessarily awkward or unrealistic. Developers ought to teach, and teachers ought to design teaching strategies. Little formal consideration has been given to the formation, training, and motivation of teacher-developer teams. Careful organization at this level is crucial.[12]

Teacher-developer teams should initially work with small groups of children selected from the larger school population on the basis of prerequisite skills and abilities related to the goals of a chosen unit of instruction. Once a pilot unit of instruction has been designed, it should be tested in one or more replications conducted by other expert teacher-developer teams on other similar groups of students. The most important requirement of these pilot replications is rapid feedback on effects so that modifications both in the teaching strategy and in the directions used to communicate it to the other teams can be made quickly and a new trial made by a new team on another group of children. When refinement of the unit has reached a plateau through this system of expert replication, it might then be implemented on a wider scale within the total school system.

The curriculum units that result from this sort of development effort would be directions to teachers for carrying out relatively short sequences of instruction. These directions would be descriptions of activities performed with success by at least one group of people who did not themselves initially develop the unit. As the instruction is successively carried out by more and more groups, the inventory of conditions under which it works successfully (and the prerequisite skills of the teachers using it) may be enlarged.

---

[12] The term *teacher* ought to refer, in this context, to *anyone* who has the responsibility for arranging the conditions under which learning actually takes place, not only the classroom teacher as traditionally conceived, but also parents, students, and other members of the community who may be called on to play teaching roles.

As the units are distributed on a successively wider scale, their usability by various groups of teachers may be systematically studied. The interaction of this kind of curriculum development with programs of teacher training, selection, and evaluation is an important area of study. Optimum curriculum development ought to be embedded in the context of a more general teacher-training and evaluation program.

Practicability

Perhaps the strongest criticism of the above strategy will be that no school system could carry it out. For the field of mathematics and science alone the technical resources and expert consultation required just to establish and maintain a viable domain-referenced objectives bank would be formidable. If other major areas of the curriculum were also included the task would obviously become impossible without extensive outside support. At best there could exist only a few such R and D systems, heavily supported by outside funds. Would they be worth establishing?

The notion of a vigorous education community, continuously and carefully assessing its needs and pursuing ways to fulfill them, is a classical and appealing one (c.f. Tyler, 1942; Chase, 1971). An attempt to maintain a few working examples, utilizing the most advanced planning, development, evaluation, and communications techniques available in the current state of the art, might be worthwhile for several reasons: (1) It might provide a counterbalance to the currently prevailing view in which the posture of school systems toward development is seen as largely defensive, with changes in curriculum and administrative organization being imported from outside in response to pressures from a chaotic variety of sources, rather than being sought out as components of an internally directed, continuous program.[13] (2) It might provide the stimulus for discovery of more effective planning, evaluation, and communications technologies. The techniques of domain-referenced testing may provide an important core of what is needed. To supplement them we particularly need a communications technology to provide convenient and effective procedures through which members of an educational community may select domains to serve as goals for individuals or groups and set priorities for development of strategies to achieve them. (3) It would provide the opportunity to study problems of research and development for whole systems as well as for development and dissemination of discrete programs or products, thereby possibly clarifying the interrelationships between the two.[14]

---

[13] See Brickell (1969) for an account of problems of assessing the effects of innovation in schools that exemplifies this largely defensive position.

Could the proposed strategy be initiated, even with substantial outside funding? The necessary technological skill could probably be assembled, and favorably disposed try-out communities might be found. The recent press for evaluation and cost-effectiveness accounting of educational programs has redirected attention to problems of assessing the activities of local school systems in detail over extended periods of time (c.f. Alkin, 1968). In the context of such local evaluation programs, together with the related pressure for performance-based teacher evaluation (McNeil, 1971; Popham, 1971b), certain school systems, perhaps working in conjunction with certain regional educational laboratories, may be more ready and able to carry out an *internal* program of research and development than would have been the case ten years ago when projects like MINNEMAST were first being supported.

Local political pressures and the network of forces that tie the typical urban school system to problems well beyond its control seem likely to undercut the proposed strategy no matter how skillful the managers. Perhaps the main problem has to do with finding ways to form more functional educational communities; communities of manageable size, linked together by certain cores of common values that could provide the basis for effective action. Perhaps in the current movement toward educational diversity, exemplified by such things as education vouchers (Center for the Study of Public Policy, 1970), ways of forming such communities may be worked out. Trying to change school systems without encouraging them to take special forms to serve the needs of special constituencies is somewhat like trying to change industrial production without allowing new corporations to be established. If and when the barriers are let down, new technologies for defining and communicating educational objectives may provide means to form these more functional educational communities.

Compromises between the internal and external models for educational research and development are obviously possible. Outside research and development organizations may borrow children and teachers from school systems for experimental purposes, under motivational conditions which may closely approximate those of idealized internal development. Outside agencies may provide access to item-form banks, assessment services, and communications schemes for schools to help them select and adapt programs (c.f. Popham, 1971c). But in any event the model of dynamic, internal development emphasizing the feed-back process by which a specific community steers its educational wagon train (Platt, 1970) ought not to be forgotten in the current focus on product development and dissemination.

---

[14] c.f. Hemphill (1969); Schutz (1970) for discussions of programs and product development.

And above the network of personal reward that produces cooperative educational development, our community must keep in focus the children. What do they know? What are they doing? What will they need to know and do in order to survive in the changing world to come?

APPENDIX 1: List of Final Units of MINNEMAST Curriculum

Grade Level	Unit Number and Title	Grade Level	Unit Number and Title
K	1. Watching and Wondering	2	15. Investigating Systems
K	2. Curves and Shapes	2	16. Numbers and Measuring
K	3. Describing and Classifying	2	17. Introducing Multiplication and Division
K	4. Using Our Senses	2	18. Scaling and Representation
K	5. Introducing Measurement	2	19. Comparing Changes
K	6. Numeration	2	20. Using Large Numbers
K	7. Introducing Symmetry	2	21. Angles and Space
1	8. Observing Properties	2	22. Parts and Pieces
1	9. Numbers and Counting	3	23. Conditions Affecting Life
1	10. Describing Locations	3	24. Change and Calculations
1	11. Introducing Addition and Subtraction	3	25. Muliplication and Motion
1	12. Measuring With Reference Units	3	26. What Are Things Made Of?
1	13. Interpretations of Addition and Subtraction	3	27. Numbers and Their Properties
1	14. Exploring Symmetrical Patterns	3	28. Mapping and Globe
		3	29. Natural Systems

APPENDIX 2: Lists of Texts, Articles, and Research Reports Produced by Staff Members of the MINNEMAST Project

a. Texts and Supplementary Materials

Douglis, A. *Ideas in Mathematics.* Philadelphia: Saunders, 1970.

Humphreys, A. H., Hindrichs, G. B., & Dadesch, R. R. *A methods manual for teaching science in elementary schools.* (Trial Version.) Minneapolis: University of Minnesota, 1964.

Humphreys, A. H., & Post, T. R. *MINNEMAST recommendations for science and mathematics in the intermediate grades.* Minneapolis: University of Minnesota, 1970.

Jones, T. B., & Ross, D. *Adventures in science and mathematics.* Minneapolis: University of Minnesota, 1970.

Rosenbloom, P. C. *Adventures with numbers.* Reading, Massachusetts: Addison-Wesley, 1958.

Rosenbloom, P. C. *Teachers manual for adventures with numbers.* Reading, Massachusetts: Addison-Wesley, 1958.

Subarsky, Z., Reed, E. W., Landin, E. R., & Klaits B. G *Living things in field and classroom.* (2nd ed.) Minneapolis: University of Minnesota, 1969.

b. Published Articles

Ahrens, R. B. MINNEMAST—The coordinated science and mathematics program. *Science and Children,* 1965, *2,* 16–18.

Bray, E. C. The MINNEMAST elementary mathematics-science program. *The Physics Teacher,* 1968, *6,* 201–206.

Cohn, A. New math—Learning by first hand experience. *Minnesota Journal of Education,* November 1963, 12–13.

Cohn, A. MINNEMAST. *Geology Teacher's Newsletter,* 1964, *1*(1), 7.

Cohn, A. Project MINNEMAST. *The Minnesotan,* November 1964, 3.

Cohn, A. Project MINNEMAST, *Alumni News,* University of Minnesota, February 1965, 11–13.

Cohn, A. What's new, math? *Twin Citian,* August 1966, 55–59.

Karplus, R. MINNEMAST science writing conference—Summer 1963. *Minnesota Journal of Science,* 1963, 7, 10–16.

Maxwell, G. Some notes and comments on Minnesota mathematics and science teaching project. *Australian Mathematics Teacher,* 1969, 25(1), 1–9.

Post, T. R. A mathematical system. *Journal of Indiana Council of Teachers of Mathematics,* Spring, 1966.

Reed, E. Instant greenhouse. *Turtox News,* 1966, *44,* 230–231.

Reed, E. Let's clean up the language. *Science and Children,* 1968, 5(5), 23–24.

Reed, E., & Ihrig, E. A. Common plants useful for classrooms. *Plant Science Bulletin,* 1968, *142,* 2–6.

Rising, G. R. Recommendations for the preparation of elementary teachers in science. *Science Education,* 1965, *49,* 359–362.

Rising, G. R. Research and development in mathematics and science education at the Minnesota school mathematics and science center and the Minnesota national laboratory. *School Science and Mathematics,* 1965, 65, 811–814.

Rosenbloom, P. C. Experimentation in 7th and 8th grade mathematics. *Minnesota Journal of Science,* 1958, *11*(2), 23–26.

Rosenbloom, P. C. Creativity in mathematics. In E. P. Torrance (Ed.), *Creative proceedings of 2nd Minnesota conference on gifted children*. Minneapolis: University of Minnesota Press, 1959, 108–112.

Rosenbloom, P. C. Implications for the colleges of the new school programs. *Mathematical Monthly*, 69, 255–259.

Rosenbloom, P. C. A leap ahead in school mathematics. In *Science and mathematics: Countdown for elementary schools*. Frontiers of Science Foundations of Oklahoma, Inc., 1960.

Rosenbloom, P. C. What is coming in elementary mathematics? *Educational Leadership*, 1960, *18*, 96–100.

Rosenbloom, P. C. Large-scale experimentation with mathematics curriculum. In Collier, R. O., & Elam, S. M., (Eds.), *Research design and analysis: Second annual Phi Delta Kappa symposium on educational research*. Bloomington: Phi Delta Kappa, 1961, 11–43.

Rosenbloom, P. C. Science instruction in high schools. *Congressional Record*, 1961, *107*(38).

Rosenbloom, P. C. The breath of school children. *The Jewish Times*, 1961, *1*(9), 5.

Rosenbloom, P. C. National conference on curriculum experimentation. *School and Society*, 1961, 89, 436–437.

Rosenbloom, P. C. Review of the child's conception of geometry by Jean Piaget. *Harvard Educational Review*, 1962, *32*, 136–141.

Rosenbloom, P. C. Mathematics K–14. *Educational Leadership*, 1962, *19*(6), 359–363.

Rosenbloom, P. C. Applied mathematics: What is needed in research and education. *SIAM Review*, 1962, *4*(4), 297–320.

Rosenbloom, P.C. Dilemma in mathematics. *Minnesota Journal of Education*, 1962, *43*, 7–8.

Rosenbloom, P. C. Minnemath science project. *Science Education News*, AAAS Misc. Pub. 62–14, December 1962, 13–14.

Rosenbloom, P. C. The Minnesota mathematics and science teaching project. *Journal of Research in Science Teaching*, 1963, *1*, 276–280.

Rosenbloom, P. C., & Hillestad, P. C. *National conference on curriculum experimentation, University of Minnesota 1961*. New York: McGraw-Hill, 1964.

Subarsky, Z. Seminar on marine biology. *The American Biology Teacher*, January 1966.

Subarsky, Z. Communication—A goal of elementary science teaching. *Science and Children*, 1966, *3*, 18–19.

Subarsky, Z. First-grade chemistry. *Science and Children*, 1966, *4*, 5–7.

Subarsky, Z. Differential heating apparatus—A model of sea and land. *Science and Children*, 1967, *4*, 4.

Subarsky, Z. Apparatus for land-water temperature study. *Science and Children*, 1967, *4*, 4.

Subarsky, Z. Apparatus for wind system demonstration. *Science and Children*, 1967, *5*, 7.

Subarsky, Z. The systems concept in science. *The Instructor*, January 1968.

Subarsky, Z. *An outline for presenting a unit on genetics*. Croft Educational Services Science/Senior High School Edition, Fall 1968.

Subarsky, Z. Curriculum construction for K–6 science and math—A strategy. *Science and Children*, 1968, *6*, 15–17.

Victor, L. Systems: An organizing principle for science curricula. *Science and Children*, 1968, *5*, 17–20.

Werntz, J. H. A style of understanding. *Nature and Science*, 1967, *4*(12).

c. Research Reports

Gottfried, N. The relationship between concepts of conservation of length and number. *Journal of Genetic Psychology,* 1969, *114,* 85–91.

Johnson, P. E. A note on methods of indexing associative relatedness. *Behavior Research Methods and Instrumentation*, March 1969.

Johnson, P. E. On the communication of concepts in science. *Journal of Educational Psychology,* 1969, *60*(1), 32–40.

Johnson, P. E. Some aspects of the psychology of written instruction. In E. Z. Rothkopt (Ed.), *Verbal learning research and the technology of written instruction*. Chicago: Rand McNally, 1969.

Johnson, P. E., & Murray, F. Some relevant and irrelevant transformations for children's concept of weight. Paper presented at the annual meeting of the American Educational Research Association, Chicago, February, 1968.

Murray, F. Reversibility training in the acquisition of length conservation. *Journal of Educational Psychology,* 1968, 59(2), 82–87.

Murray, F. Operational conservation of illusion-distorted length. *British Journal of Educational Psychology,* 1968, 38, 189–193.

Murray, F. Phenomenal-real discrimination and the conservation of illusion-distorted length. *Canadian Journal of Psychology,* 1968, *22*(2), 114–121.

Victor, L. The development of modern space-time concepts in the elementary schools. *Journal of Research in Science Teaching,* 1969, *6*(1), 36–41.

Victor, L. Conceptual schemes of science: A call for research. *Science Education*, 1969, *53*(4), 335–339.

APPENDIX 3: Examples of Item Forms Developed by Members of the Evaluation Staff of the MINNEMAST Project

Item Form	Title
2.2	Producing examples of simple and non-simple, open and closed curves.
3.6	Using a comparison or "control" set to detect changes in a corresponding set.
3.15	Comparing numerosity of sets by one-to-one correspondence.
9.7	Producing a number satisfying a given relation to specified number(s) (spoken form).
9.8	Producing a number satisfying a given relation to specified number(s) (written form).
16.14	Comparing two objects on equal-arm balance and choosing symbol to complete statement of the weight relation.
17.8	Completing multiplication equation by writing the product of two factors.
23.10	Explaining how to set up an experiment to study moisture effects on sowbugs and mealworm beetles.
26.2	Plotting a single point on a volume-weight graph.

ITEM FORM 2.2*

Producing examples of simple and non-simple, open and closed curves.

GENERAL DESCRIPTION

The child is given an example of a simple open, simple closed, non-simple open, or non-simple closed curve and asked to draw several more that are different, but of the same kind.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

Child is given an example of the required type of curve at the beginning. Child produces curves by drawing them.

Distinguishing Among Cells

Type of curve required: (1) simple open, (2) simple closed, (3) non-simple open, (4) non-simple closed. (The last two curve types are not standard topological classifications, but are clearly defined.)

Varying Within Cells

Instances of sample curves presented.

CELL MATRIX

Script (b)

Simple closed	(1)
Simple open	(2)
Non-simple closed	(3)
Non-simple open	(4)

(Sample curve is drawn from replacement set corresponding to script.)

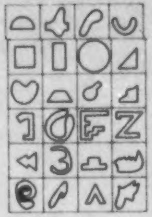
ITEM FORM SHELL

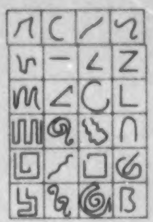
MATERIALS Curve card (a) Response Sheet Pencil	**ITEM FORM: 2.2** **CELL:** 1 **REPLICATION:** 1
DIRECTIONS TO E	**SCRIPT**
Don't look at curve card yourself, until you have laid it in front of S.	Here is a (b) simple closed curve. Here is a pencil and paper. Draw another (b) simple closed curve that is different from that one. (Answer #1)
After S finishes each answer, write its number beside it. If you aren't sure whether S is finished, ask him.	
	Now, draw another (b) simple closed curve that is different. (Answer #2)
In transition to each new question, you can say "um hum" or "O.K." but don't say "good" or otherwise put special emphasis on correct answers.	
	Now, draw another (b) simple closed curve (Answer #3)
	Now draw another (b) simple closed curve that is different. (Answer #4)
RECORDING Attach response sheet.	

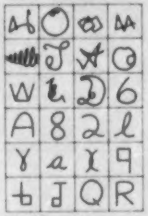
REPLACEMENT SCHEME

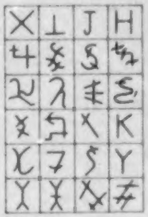
Curve Cards (a)	Script (b)
Cell 1: choose from R.S. 2.1.	Cell 1: simple closed
Cell 2: choose from R.S. 2.2.	Cell 2: simple open
Cell 3: choose from R.S. 2.3.	Cell 3: non-simple closed
Cell 4: choose from R.S. 2.4.	Cell 4: non-simple open

REPLACEMENT SETS


R.S. 2.1.
Simple, closed curves


R.S. 2.2.
Simple, open curves


R.S. 2.3.
Non-simple, closed curves


R.S. 2.4.
Non-simple, open curves

SCORING SPECIFICATIONS

Cell 1 (simple closed): Curve bounds an area, may not have crossing points.

Cell 2 (simple open): Curve does not bound an area, may not have crossing points.

Cell 3 (non-simple closed): Part of curve bounds an area, must have at least one crossing point.

Cell 4 (non-simple open): No part of curve bounds an area, must have at least one crossing point.

---

* Originally developed by Stephen Lundin.

ITEM FORM 3.6*

Using a comparison or "control" set to detect changes in a corresponding set.

GENERAL DESCRIPTION

The child is given two approximately identical sets of objects and asked to establish a one-to-one correspondence between them. One set (the "experimental" set) is then covered and one of its members may or may not be replaced by another one. The child is then asked whether or not any of the properties of the objects in the set have changed.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for all Cells

Objects are abstract "property cards" (RS 3.6.1).

Distinguishing among Cells

Number of objects in sets (either five or eight). Whether a change is, or is not made in the "experimental" set.

Varying Within Cells

Instances of property cards chosen to make up sets. Instances of property card chosen as replacement for a member of the "experimental" set.

CELL MATRIX

	Change	No Change
Sets of five members	(1)	(2)
Sets of eight members	(3)	(4)

---

* Originally developed by Bruce Mussell.

ITEM FORM SHELL

MATERIALS	
1. Set of cards (a) 2. Set of cards (b)	
DIRECTIONS TO E	**SCRIPT**
Place cards in Set 1 in a horizontal line at the edge of the board closest to S. Place cards in Set 2 in a horizontal line directly above Set 1.	
Then say:	Show me if these two sets are alike.
If S does not show a 1–1 correspondence, say:	Can you pair these objects by similar properties?
If S does not respond, STOP.	
If S says "yes," say:	Do it please.
When S has correctly paired the objects leave Set 1 as the child has arranged it.	
Move set 2 closest to you, cover the set and say:	I'm going to cover this set for a moment.
Take card (c), put hand with object under covering. Rearrange set and (d). Then say:	Now, have any of the properties of the objects in this set changed?
Regardless of response, say:	Show me.
Keep a running record of what S does and says.	

REPLACEMENT SCHEME

Sets of Objects (a), (b)

Cells 1 and 2: Choose five ordered pairs from R.S. 3.6.1., assign the first member to (a) and the second member to (b)

Cells 3 and 4: Take all eight ordered pairs from R.S. 3.6.1., assign the first member to (a) and the second member to (b)

Substitution Object (c)

All cells: Choose an object from R.S. 3.6.2. such that it is not a member of sets (a) or (b).

Directions to E (d)

Cells 1 and 3: Remove an object from the set and replace it with card (c). Record which card you removed.

Cells 2 and 4: Act in the same way as when replacing a card but do not change cards. Just put card (c) in and then remove it.

REPLACEMENT SETS

R.S. 3.6.1 (Numbers refer to property cards identified in R.S. 3.6.2): (1, 46), (2, 47), (4, 48), (5, 49), (10, 50), (11, 51), (13, 52), (14, 53).

R.S. 3.6.2 (Property cards):

	RED	BLUE	YELLOW
SQUARE	spotted 1,46 striped 2,47 plain 3	spotted 4,48 striped 5,49 plain 6	spotted 7 striped 8 plain 9
CIRCLE	spotted 10,50 striped 11,51 plain 12	spotted 13,52 striped 14,53 plain 15	spotted 16 striped 17 plain 18
TRIANGLE	spotted 19 striped 20 plain 21	spotted 22 striped 23 plain 24	spotted 25 striped 26 plain 27

NOTE: Object 46 is exactly like object 1, object 47 is exactly like object 2, etc.

SCORING SPECIFICATIONS

Child should state correctly (yes or no) whether any properties of the set had changed and support his answer by pointing to the changed card.

ITEM FORM 3.15*

Comparing numerosity of sets by one-to-one correspondence.

GENERAL DESCRIPTION

The child is given either two or three sets of "counters," each having approximately 20 members (or less). The sets may have the same number of members or they may differ by one member. The child is asked to show whether or not the sets have the same number of members, without counting.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for all Cells

Only standard "counters" (small colored disks) are used. Each set of counters is a different color (red, green or yellow).

Distinguishing between cells

Number of sets compared (two or three). Whether or not the sets have the same number of objects. Approximate number of objects in each set (about 5, about 13, or about 21).

Varying with Cells

No variation.

CELL MATRIX

Approximate Number of Objects in Sets	Number of Sets Compared				
	2		3		
	Equality Relations		Equality Relation		
	$N_a = N_b$	$N_a \neq N_b$	$N_a = N_b = N_c$	$N_a = N_b \neq N_c$	$N_a \neq N_b \neq N_c$
5	(1)	(4)	(7)	(10)	(13)
13	(2)	(5)	(8)	(11)	(14)
21	(3)	(6)	(9)	(12)	(15)

---

* Originally developed by Bruce Mussell.

ITEM FORM SHELL

MATERIALS 1. Set of counters (a) 2. Set of counters (b) 3. Set of counters (c)	
DIRECTIONS TO E	SCRIPT
Place the above sets near either edge (and the middle) of the test board as shown above. Then say:	Show me if these (d) sets (point) have the same number of members.
If S begins to count or says "I don't know how," say:	In class you paired objects to tell if two sets had the same number of members. Please show me if these (d) sets have the same number of members.
When S has finished, say:	Do they have the same number of members?
Keep a running record of what S does and says.	

REPLACEMENT SCHEME

Sets of Counters (a) = red, (b) = green, (c) = yellow

Number of objects in each set:

Cell 1: (a) 5; (b) 5
Cell 2: (a) 13; (b) 13
Cell 3: (a) 21; (b) 21
Cell 4: (a) 5; (b) 6
Cell 5: (a) 13; (b) 14
Cell 6: (a) 21; (b) 22
Cell 7: (a) 5; (b) 5; (c) 5
Cell 8: (a) 13; (b) 13; (c) 13
Cell 9: (a) 21; (b) 21; (c) 21
Cell 10: (a) 5; (b) 5; (c) 6
Cell 11: (a) 13; (b) 13; (c) 14
Cell 12: (a) 21; (b) 21; (c) 22
Cell 13: (a) 5; (b) 6; (c) 7
Cell 14: (a) 13; (b) 14; (c) 15
Cell 15: (a) 21; (b) 22; (c) 23

Script (d):

Cells 1 through 6: "two"
Cells 7 through 15: "three"

SCORING SPECIFICATIONS

Child should state correctly (yes or no) whether or not the sets have the same number of members. He should also carry out a detectable one-to-one pairing operation.

ITEM FORM 9.7*

Producing a number satisfying a given order relation to specified numbers(s) (spoken form).

GENERAL DESCRIPTION

The child is asked to say the name of a number that bears a specified order relation ("greater than" or "less than") to a given number or numbers in the range 0 through 20. Given numbers are presented in spoken form and response is spoken.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The presentation is completely spoken; a spoken response is required.

Distinguishing Among Cells

Three scripts are used asking respectively for a number greater than a given number, for a number less than a given number, and for a number greater than one given number and less than another.

Within the third script, three conditions are allowed: (1) first given numeral greater than second with required number possibly an integer; (2) first given numeral greater than second with required number necessarily not an integer; and (3) first given numeral less than second so that the solution to the problem is the empty set.

Varying Within Cells

Within each cell, the given numbers are integers from the range 0 through 20 chosen so that the correct response (when it is not the empty set) can be a real number from the range 0 through 20.

CELL MATRIX

Script (a)	"greater than b_1"	"less than b_1"	"greater than b_1" but less than b_2"		
Numerals (b)	$0 \leqq b_1 \leqq 19$	$1 \leqq b_1 \leqq 20$	$0 \leqq b_1 \leqq 18$ $b_1 + 2 \leqq b_2 \leqq 20$	$0 \leqq b_1 \leqq 19$ $b_2 = b_1 + 1$	$1 \leqq b_1 \leqq 20$ $0 \leqq b_2 < b_1$
	(1)	(2)	(3)	(4)	(5)

ITEM FORM SHELL

MATERIALS None	
DIRECTIONS TO E Read script to child. Write down child's exact words.	SCRIPT Tell me a number that is __________

REPLACEMENT SCHEME

(a) Script

Cell 1: "less than b_1" "greater than b_1."
Cells 3,4,5: "greater than b_1 but less than b_2."

(b) Numerals within Script

Cell 1: Choose b_1 from R.S. 9.1

Cell 2: Choose b_1 from R.S. 9.2

Cell 3: Choose two numbers from R.S. 9.3

Cell 3: Choose two numbers from R.S. 9.3
Let b_1 = smaller number; b_2 = larger number
Reject if $b_2 - b_1 \leq 1$

Cell 4: Choose b_1 from R.S. 9.3
Let $b_2 = b_1 + 1$

Cell 5: Choose two numbers from R.S. 9.3
Let b_1 = larger number; b_2 = smaller number
Reject if $b_1 = b_2$

REPLACEMENT SETS

R.S. 9.1: Whole numbers 0,1,2, . . . ,19.

R.S. 9.2: Whole numbers 1,2,3, . . . ,20.

R.S. 9.3: Whole numbers 0,1,2, . . . ,20.

SCORING SPECIFICATIONS

Cell 1: Any real number x where $x > b_1$

Cell 2: Any real number x where $x < b_1$

Cell 3: Any real number x where $b_1 < x < b_2$

Cell 4: Any real number x where $b_1 < x < b_2$

Cell 5: Any response equivalent to saying that there are no numbers which can fulfill the conditions.

---

* Originally developed by Donald Sension.

ITEM FORM 9.8*

Producing a number satisfying a given order relation to specified number(s) (written form).

GENERAL DESCRIPTION

The child is asked to write down a number that bears a specified order relation ("greater than" or "less than") to a given number or numbers in the range 0 through 20. Given numbers are presented in written form and response is written.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

Numerals are presented in written form, together with spoken instructions. A written response is required.

Distinguishing Among Cells

Same as for Item Form 9.7

Varying within Cells

Same as for Item Form 9.7

CELL MATRIX

Same as for Item Form 9.7

ITEM FORM SHELL

MATERIALS Pencil Response Sheet Numeral card(s) (b)	
DIRECTIONS TO E Place materials in front of child. Point appropriately to numeral card(s) as you say: When child has finished, attach Response Sheet to this page.	**SCRIPT** Write a number that is_____ ____________________

DESCRIPTION OF MATERIALS

Set of Numeral Cards (T.O.S. 9.5.0): Cards, approximately 4" × 6", each displaying one of the numerals b where $b \in \{0, 1, 2, \ldots, 20\}$ in ¼" high lettering.

REPLACEMENT SCHEME

(a) Script
  Cell 1: "greater than this one."
  Cell 2: "less than this one."
  Cell 3: "greater than this one but less than this one."
(b) Numeral (on Numeral card)
  Same as for Item Form 9.7

REPLACEMENT SETS

Same as for Item Form 9.7

SCORING SPECIFICATIONS

Same as for Item Form 9.7 except that response is written.

---

* Originally developed by Donald Sension.

ITEM FORM 16.14*

Comparing two objects on equal-arm balance and choosing a symbol to complete a statement of the weight relation.

GENERAL DESCRIPTION

The child is asked to compare the weights of two objects that may be (1) indistinguishable by hefting but easily distinguished on the balance, (2) indistinguishable even on the balance. In each of these situations, size varies as an irrelevant dimension. An equal-arm balance is available but instructions for its use are non-directive. The child is asked to select one of the three symbols (>, <, and =) and place it in the blank space provided between the two weight symbols.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The equal-arm balance is of similar construction to that used in MINNEMAST Unit 16, made of Tinkertoys, cardboard, string, a metal weight, and a foot ruler.

The objects are opaque, cylindrical bottles, identical except for weight (either 23 gm. or 25 gm.) and size (either 2" × ⅝" or 2½" × 1⅜"). Each is identified by a lower-case letter assigned at random.

The child is asked to complete a symbolic statement, corresponding to the weight relation, by choosing the correct relation symbol.

Distinguishing among Cells

Three weight relations (detectable by balance only, not by hefting or "feel") defined in terms of the location of the objects when placed in front of the child:

left > right; left < right; left = right.

Three size relations:

left > right; left < right; left = right.

CELL MATRIX

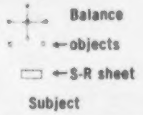
	Weight Relations (Detectable by Balance Only)		
Size Relations	$W_1 > W$	$W_1 < W_r$	$W_1 = W_r$
$S_1 > S_r$	(1)	(4)	(7)
$S_1 < S_r$	(2)	(5)	(8)
$S_1 = S_r$	(3)	(6)	(9)

* Originally developed by Wells Hively.

ITEM FORM SHELL

<table>
<tr><td>MATERIALS

Beam Balance
Objects 1 and r
from T.O. 16.14.0
Stimulus-Response sheet
(attached)
Pencil</td><td></td></tr>
<tr><td>DIRECTIONS TO E

Place materials in front of child. (Keep order of objects given above.)
 </td><td>SCRIPT

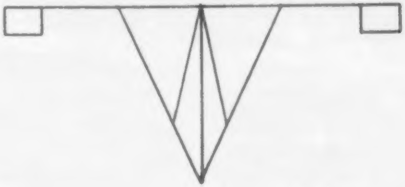
Here are two objects. They have symbols attached to them. Compare them by weight and write one of these three signs (point) in the blank (point) to form the comparison sentence.

You may use this balance if you need to.</td></tr>
</table>

RECORDING

Attach Stimulus-Response sheet to this page.
Describe what child did.

If balance was used, insert object symbols in schematic drawing of the balance given below, and mark the position of the plumb-line at the time of child's judgment.



DESCRIPTION OF MATERIALS

Pencil (T.O. 16.1.1)

Beam Balance (T.O. 16.13.1): Equal-arm beam balance made from tinker-toy materials as decribed in MINNEMAST Unit 16.

Set of Weight Comparison Objects (T.O. 16.14.0): Set of opaque plastic cylindrical bottles with firmly fitting lids. Two sizes of bottles have been chosen. The small bottle has a length of 2" and a diameter of ⅝". The large bottle has a length of 2½" and a diameter of 1⅜". Two weight values have been chosen so that the objects cannot typically be distinguished by hefting but can be distinguished on the balance. Each object is designated by a randomly chosen, lower-case letter.

	Weight	
Size	23 gm	25 gm
small	a	m k
large	b	o n

Stimulus-Response sheet (attached to item) (T.O. 16.14.1): a sheet of paper approximately 6" × 4" with the following display:

> Write >, <, or = in the blank
>
> W_1____________W_r

where 1 and r are the appropriate subscripts (from Replacement Scheme).

REPLACEMENT SCHEME

(l,r) Objects

Cell 1:	(o,a)	
Cell 2:	(m,b)	
Cell 3:	Choose	from R.S. 16:13
Cell 4:	(b,m)	
Cell 5:	(a,o)	
Cell 6:	Choose	from R.S. 16.14
Cell 7:	Choose	from R.S. 16.15
Cell 8:	Choose	from R.S. 16.16
Cell 9:	Choose	from R.S. 16.17

REPLACEMENTS SETS

R.S. 16.13	Ordered pairs	(m,a);	(o,b)
R.S. 16.14	Ordered pairs	(a,m);	(b,o)
R.S. 16.15	Ordered pairs	(b,a);	(o,m)
R.S. 16.16	Ordered pairs	(a,b);	(m,o)
R.S. 16.17	Ordered pairs	(m,k);	(o,n)

SCORING SPECIFICATIONS

A correct response is made by writing the correct symbol (>, <, or =) in the blank space to complete the comparison sentence. This should be > in Cells 1, 2, and 3; < in Cells 4, 5, and 6; = in Cells 7, 8, and 9.

ITEM FORM 17.8*

Completing a multiplication equation by writing the product of two factors.

GENERAL DESCRIPTION

The child is presented with an incomplete multiplication equation (missing product) and is asked to complete it. One of three manipulanda (a number line, a set of parallel number lines, or a set of counters) is made available with non-directive instructions for its use.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The S–R sheet contains a written incomplete multiplication equation (missing product).

Distinguishing Among Cells

Equation Format (two levels); (i) product on left; (ii) product on right.

Second Factor Values (two levels); (i) values same as those used in Unit; (ii) extension beyond values used in Unit.

Script (four levels): using vocabulary "problem," "equation," "multiplication," "product."

Manipulanda (three levels); (i) number line: (ii) two parallel number lines; (iii) counters.

Varying Within Cells

First factor: one of the numerals 1, 2, 3, 4, 5, 6.

Second factor, level 1: one of the numerals 1, 2, 3, 4, 5, 6.

Second Factor, level 2: one of the numerals 7, 8, 9, 10.

CELL MATRIX

Equation Format (A)	Second Factor (B)	Script (C)	I	II	III
1	1	1	1	17	33
		2	2	18	34
		3	3	19	35
		4	4	20	36
	2	1	5	21	37
		2	6	22	38
		3	7	23	39
		4	8	24	40
2	1	1	9	25	41
		2	10	26	42
		3	11	27	43
		4	12	28	44
	2	1	13	29	45
		2	14	30	46
		3	15	31	47
		4	16	32	48

ITEM FORM SHELL

MATERIALS Pencil S–R sheet (attached) (b) ____________________	
DIRECTIONS TO E Place pencil and S–R sheet on desk in front of child. Point to equation and say: Place______(b)____on desk in front of child as you say: RECORDING: Attach S–R sheet to this page. Record all verbal responses the child makes. Describe (or depict) how the child used manipulanda.	SCRIPT (c) ____________________ Write your answer in this blank. You may use this if you need to.

DESCRIPTION OF MATERIALS

Pencil (T.O. 17.1.1):

S–R sheet (attached to item) (T.O. 17.16.1): sheet of paper, about 4" × 6" displaying equation format____________________(a)____________________.

Number Line (T.O. 17.5.1): Drawing of a number line, 20½ inches long, marked in integers from 0 through 65.

Parallel Number Lines (T.O. 17.8.1): Drawing of a number line, 20½ inches long, marked in integers from 0 through 65. Parallel to this first line a second line the same except for the absence of the integers.

Counters (T.O. 17.9.1): Set of small, cardboard counters.

REPLACEMENT SCHEME

(a) Equation Format (on S–R sheet)

Cells 1–8, 17–24, 33–40: A × B = ____
Cells 9–16, 25–32, 41–48: ____ = A × B

(b) Manipulanda

Cells 1–16:	Number Line
Cells 17–32:	Parallel Number Lines
Cells 33–48:	Counters

(A) First Factor

All Cells: Choose from R. S. 17.1

(B) Second Factor

Cells 5–8, 13–16, 21–24, 29–32, 37–40, 45–48: Choose from R. S. 17.1
Cells 1–4, 9–12, 17–20, 25–28, 33–36, 41–44: Choose from R.S. 17.6

(C) Script

Every 4th cell (i.e. 1, 5, 9, . . ., 41, 45): "Solve this problem."
Every 4th cell (i.e. 2, 6, 10, . . ., 42, 46): "Complete this equation."
Every 4th cell (i.e. 3, 7, 11, . . ., 43, 47): "Complete this multiplication equation."
Every 4th cell (i.e. 4, 8, 12, . . ., 44, 48): "Find the product."

REPLACEMENT SETS

R.S. 17.1	Set of numerals: 1, 2, 3, 4, 5, 6
R.S. 17.6	Set of numerals: 7, 8, 9, 10

SCORING SPECIFICATIONS

A correct response is made by writing in the blank space of the equation format the correct numeral for the product of the two given factors.

---

* Originally developed by Eugene Lenarz.

ITEM FORM 23.10*

Explaining how to set up an experiment to study moisture effects on sowbugs and mealworm beetles.

GENERAL DESCRIPTION

The child is given materials necessary to carry out a simple experiment and asked how to use them to set up an experiment to study how moisture affects a sowbug or a mealworm beetle.

STIMULUS AND RESPONSE CHARACTERISTICS

There are only two items in this item form, one per cell. The same script applies to each but different materials are presented.

CELL MATRIX

Materials	Cell
First Set (petri dish, etc.)	1
Second Set (plastic tray, etc.)	2

* Originally developed by Bruce Mussell.

ITEM FORM SHELL

MATERIALS (a)	
DIRECTIONS TO E Place the materials on the table in front of the child randomly and say: If no response, say:	SCRIPT Here are some materials you can use. Suppose we want to study how moisture affects a sowbug or a mealworm beetle. How could we set up an expirement to do this? Moisture means wetness.
RECORDING Write down child's exact words and describe what he does with the materials.	

DESCRIPTION OF MATERIALS

Set 1 (T.O.S. 23.9.0): Petri dish, piece of blotter cut to size of petri dish, scotch tape, scissors, beaker of water, medicine dropper.

Set 2 (T.O.S. 23.10.0): Rectangle plastic tray top, piece of cloth about one inch shorter and one inch narrower than the tray top, glue, scissors, beaker of water, straw.

REPLACEMENT SCHEME

(a) Set of Materials
  Cell 1: Set 1
  Cell 2: Set 2

SCORING SPECIFICATIONS

Rather than score the responses "correct/incorrect," classify them into general categories according to their completeness and comprehensibility and the provision for studying both wet and dry conditions.

ITEM FORM 26.2*

Plotting a single point on a volume-weight graph.

GENERAL DESCRIPTION

A graph, with axes indicating volume and weight, and a sheet displaying either an ordered pair or a volume-weight chart is presented. The child is asked to plot the point represented by the data onto the grid.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The grid has the characteristics described in the Description of Materials.

Distinguishing Among Cells

The child is given the data either as an ordered pair or as a Volume-Weight chart.

The data are such that the point to be plotted is either at the intersection of two grid lines, or on an X-axis grid line at a position intermediate (in tenths) between two Y-axis grid lines.

Complete crossing of these categories yields four cells.

Varying Within Cells

The data for the point to be plotted are varied within the limits of the grid and of the Cell Constants specifications.

For Cells 1 and 2, the Volume and Weight values are both chosen from the set of integers 1 through 12, with the requirement that the two values must not be identical. (This condition eliminates situations where order would not matter.)

For Cells 3 and 4, the Volume value is chosen from the set of integers 1 through 12; and the Weight value, $j^{k}{}_{/10}$ units is chosen so that j is from the set of integers 0 through 11, and k is from the set of integers 1 through 9.

CELL MATRIX

	Y-coordinate an integer	Y-coordinate in tenths
Date as Ordered pair	(1)	(3)
Data as V/W Chart	(2)	(4)

---

* Originally developed by Graham Maxwell.

ITEM FORM SHELL

MATERIALS Stimulus Sheet (attached) Grid (attached) Pencil	
DIRECTIONS TO E Place materials in front of child and point to the relevant parts as you say: When child has finished, attach both the stimulus sheet and the grid to this page.	SCRIPT (d)________________

DESCRIPTION OF MATERIALS

Stimulus Sheet (attach one of the following objects to the item as specified by (a) in the Replacements).

(T.O. 26.5.1): A sheet of 6"×4" notepaper displaying the ordered pair P [(b), (c)];

(T.O. 26.4.1): A sheet of 6"×4" notepaper displaying the following labeled chart:

OBJECT	VOLUME (in units of volume)	WEIGHT (in units of weight)
P	(b)	(c)

Grid (attached to item) (T.O. 26.2.1): A sheet of paper displaying a grid, 6" × 6", with grid lines ½" apart. On each axis, the grid lines are marked with the numbers 1 through 12. The X-axis is labeled "Volume (in units of volume)," and the Y-axis is labeled "Weight (in units of weight)."

Pencil (T.O. 26.1.1):

REPLACEMENT SCHEME

(a) Stimulus Sheet

Cells 1 and 3:	T.O. 26.5.1
Cells 2 and 4:	T.O. 26.4.1

(b,c) Coordinates of point P for Stimulus Sheet

Cells 1 and 2:	Choose b	from R.S. 26.1
	Choose c	from R.S. 26.1
	Reject if $b = c$	
Cells 3 and 4	Let $b = i$	
	$c = j^{k}{}_{/10}$	
	choose i	from R.S. 26.1
	choose j	from R.S. 26.2
	choose k	from R.S. 26.3

APPENDIX 4: Examples of Sections of Reports on selected MINNEMAST Units

Example	Description
1	Extract from the summary of Results of Testing Unit 9: Numbers and Counting, Fall 1968.
2	Extract from the Report on Results of Testing Unit 16: Numbers and Measuring, Feb.-Mar.-Apr. 1969.
3	Extract from the Report on Results of Testing Unit 16: Numbers and Measuring, Feb.-Mar.-Apr. 1969.
4	Extract from the Report on Results of Testing Unit 23: Conditions Affecting Life, Spring 1969.

APPENDIX 4, Example 1

Extract from the Summary of Results of Testing Unit 9:
Numbers and Counting, Fall 1968*

Table showing proportions of correct responses for the cells of ITEM FORMS 9.7 and 9.8: Producing Number (x) satisfying given relation to specified numbers (b_1 and b_2)**

		Response Requirement				
Item Form	Response Form	(1) $x > b_1$	(2) $x < b_1$	(3) $b_1 < x < b_2$ $b_2 > b_1 + 1$	(4) $b_1 < x < b_2$ $b_2 = b_1 + 1$	(5) $b_1 < x < b_2$ $b_2 \quad b_1$
9.7	Spoken	$\frac{18}{20}$	$\frac{9}{10}$	$\frac{16}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
9.8	Written	$\frac{18}{20}$	$\frac{10}{10}$	$\frac{11}{20}$	$\frac{0}{20}$	$\frac{0}{20}$
	Totals:	$\frac{36}{40}$ (.90)	$\frac{19}{20}$ (.95)	$\frac{27}{40}$ (.68)	$\frac{1}{40}$ (.03)	$\frac{2}{40}$ (.05)

---

* Prepared by Donald Sension.

** See Appendix 3 for the complete Item Forms.

APPENDIX 4, Example 2

Extract from the Report on Results of Testing Unit 16:
Numbers and Measuring, Feb.-Mar.-Apr. 1969*

Item Form 16.6: Understanding T-Notation

Description: A three-place numeral was presented in T-notation.
The verbal instruction was:
"Write down the numeral indicated by this T-notation."

Summary of test Results

Cell 1: Presented $i\,T \cdot T + 0\,T + K =$ _______
(where i and k are non-equal digits from the range 1 thru 9.)

Completely satisfactory	3	(12%)
Unsatisfactory		
ik	5	
i + k	4	
other	11	
No response	2	
	25	

Cell 2: Presented i T·T + K = ———
(where i and k are non-equal digits from the range 1 thru 9)

Completely satisfactory	1	(4%)
Unsatisfactory		
ik	8	
ki	1	
l + k	3	
other	8	
No response	4	
	25	

Cell 3: Presented i T·T + j T + 0 = ———
(where i and j are non-equal digits from the range 1 thru 9)

Completely satisfactory	4	(16%)
Unsatisfactory		
ij	2	
i + j	6	
other	11	
No response	2	
	25	

Cell 4: Presented i T·T + j T = ———
(where i and j are non-equal digits from the range 1 thru 9)

Completely satisfactory	1	(4%)
Unsatisfactory		
ij	4	
i + j	2	
(i + j)0	3	
other	12	
No response	3	
	25	

Cell 5: Presented i T·T + j T + k = ———
(where i, j and k are non-identical digits from the range 1 thru 9)

Completely satisfactory	1	(4%)
Unsatisfactory		
l + j + k	3	
other	13	
No response	8	
	25	

---

* Prepared by Graham Maxwell.

Listing of Test Results: Item Form Cell 16.6.1

Class	Child	Stimulus	Response	Score*
30	6602897	9 TT + 0 T + 5	905	CS
42	6602925	4 TT + 0 T + 2	402	CS
11	6600262	1 TT + 0 T + 2	102	CS
64	6600287	9 TT + 0 T + 8	98	UN
64	6602943	7 TT + 0 T + 2	72	UN
68	6602957	4 TT + 0 T + 3	43	UN
37	6600349	2 TT + 0 T + 5	25	UN
69	6600084	3 TT + 0 T + 8	38	UN
69	6600075	5 TT + 0 T + 1	6	UN
62	6600869	1 TT + 0 T + 3	4	UN
30	6600186	8 TT + 0 T + 4	12	UN
13	6600508	2 TT + 0 T + 1	3	UN
37	6600504	5 TT + 0 T + 6	16	UN
11	6600866	6 TT + 0 T + 7	10	UN
13	6600162	9 TT + 0 T + 6	9	UN
12	6602875	8 TT + 0 T + 4	42	UN
12	6602874	4 TT + 0 T + 9	"I don't get what you're talking about." ERQ. 12	UN
65	6600278	8 TT + 0 T + 2	22	UN
62	6600632	5 TT + 0 T + 7	15	UN
38	6600503	5 TT + 0 T + 3	27	UN
40	6600498	4 TT + 0 T + 1	4	UN
28	6600255	7 TT + 0 T + 4	41	UN
63	6600246	8 TT + 0 T + 2	39	UN
47	6600605	6 TT + 0 T + 2	No Response	NR
65	6600222	4 TT + 0 T + 7	No Response	NR

*Legend: CS = Completely Satisfactory; UN = Unsatisfactory; NR = No Response.

Listing of Test Results (cont'd): Item Form Cell 16.6.2

Class	Child	Stimulus	Response	Score*
65	6600279	6 TT + 4	604	CS
30	6600509	6 TT + 5	65	UN
40	6600493	9 TT + 4	94	UN
36	6600342	7 TT + 9	79	UN
37	6600197	9 TT + 3	93	UN
42	6602914	1 TT + 6	16	UN
11	6600619	6 TT + 7	67	UN
36	6600187	2 TT + 8	28	UN
28	6600850	2 TT + 9	29	UN
62	6600623	4 TT + 5	54	UN
69	6602967	9 TT + 2	11. "Would I add 9 tens and 2 tens?"	UN
69	6600076	7 TT + 1	8	UN
68	6600077	2 TT + 1	3	UN
13	6602679	4 TT + 3	"Okay that would be 400. Does this T go to the four and this to the three?"	UN
42	6602927	8 TT + 1	90	UN
65	6600627	1 TT + 7	1	UN
64	6600243	4 TT + 7	20	UN
28	6600255	5 TT + 4	29	UN
12	6602304	3 TT + 9	100	UN
37	6602906	3 TT + 1	81	UN
47	6602938	3 TT + 6		
47	6600586	3 TT + 7	"I don't think I can do it"	NR
40	6600195	8 TT + 4	No Response	NR
13	6602883	7 TT + 4	No Response	NR
63	6602751	5 TT + 9	No Response	NR

* Legend: CS = Completely Satisfactory; UN = Unsatisfactory; NR = No Response.

Listing of Test Result (cont'd): Item Form Cell 16.6.3

Class	Child	Stimulus	Response	Score*
11	6602858	6 TT + 5 T + 0	650	CS
30	6602900	2 TT + 9 T + 0	290	CS
42	6602928	3 TT + 4 T + 0	340	CS
11	6600258	9 TT + 1 T + 0	910	CS
59	6600083	4 TT + 5 T + 0	45	UN
64	6600211	3 TT + 7 T + 0	37	UN
40	6600205	5 TT + 9 T + 0	14	UN
42	6602922	6 TT + 9 T + 0	15	UN
62	6600252	6 TT + 2 T + 0	8	UN
47	6602939	5 TT + 5 T + 0	10	UN
62	6600251	5 TT + 8 T + 0	13	UN
36	6600499	5 TT + 2 T + 0	7	UN
63	6600254	5 TT + 7 T + 0	10	UN
40	6600501	7 TT + 4 T + 0	101	UN
64	6600280	3 TT + 1 T + 0	70	UN
28	6600217	3 TT + 5 T + 0	10	10
13	6604004	7 TT + 5 T + 0	"You mear copy it?" ERQ	UN
68	6600098	5 TT + 6 T + 0	41	UN
65	6600272	1 TT + 3 T + 0	111	UN
12	6602865	7 TT + 4 T + 0	41	UN
63	6600854	8 TT + 9 T + 0	180	UN
28	6600872	7 TT + 6 T + 0	130	UN
37	6600207	7 TT + 9 T + 0	0	UN
36	6600441	2 TT + 3 T + 0	No Response	NR
30	6600437	5 TT + 8 T + 0	No Response	NR

* Legend: CS = Completely Satisfactory; UN = Unsatisfactory; NR = No Response.

Listing of Test Results (cont'd): Item Form Cell 16.6.4

Class	Child	Stimulus	Response	Score*
47	6600596	4 TT + 6 T	460	CS
11	6600868	4 TT + 9 T	49	UN
11	6600210	7 TT + 4 T	74	UN
37	6600344	3 TT + 7 T	37	UN
68	6600826	1 TT + 6 T	16	UN
12	6604003	4 TT + 1 T	5	UN
40	6600205	4 TT + 8 T	12	UN
37	6600428	7 TT + 8 T	150	UN
40	6602911	1 TT + 6 T	70	UN
65	6600234	4 TT + 3 T	70	UN
13	6600162	3 TT + 8 T	3	UN
62	6600285	9 TT + 8 T	19	UN
63	6600852	6 TT + 9 T	28	UN
42	6602919	1 TT + 9 T	110	UN
12	6602875	5 TT + 7 T	42	UN
69	6600078	9 TT + 7 T	11	UN
13	6604005	7 TT + 9 T	"It can't be 79 can it?" 19	UN
30	6600500	9 TT + 6 T	14	UN
28	6602887	5 TT + 6 T	36	UN
68	6600093	6 TT + 5 T	120	UN
63	6602942	8 TT + 9 T	1	UN
42	6602926	2 TT + 8 T	38	UN
64	6602945	3 TT + 5 T	"I don't get it."	NR
36	6600196	6 TT + 7 T	No Response	NR
28	6600227	4 TT + 2 T	No Response	NR

* Legend: CS = Completely Satisfactory; UN = Unsatisfatcory; NR = No Response.

Listing of Test Results (cont'd.). Item Form Cell 16.6.5

Class	Child	Stimulus	Response	Score*
42	6602920	1 TT + 5 T + 9	159	CS
42	6602918	5 TT + 6 T + 1	12	UN
69	6600075	6 TT + 2 T + 7	15	UN
47	6602939	6 TT + 1 T + 8	15	UN
63	6600854	9 TT + 2 T + 8	20	UN
47	6600601	3 TT + 4 T + 1	10	UN
64	6600857	7 TT + 2 T + 6	35	UN
63	6600860	1 TT + 9 T + 2	294	UN
65	6600851	1 TT + 6 T + 5	20	UN
12	6602868	6 TT + 8 T + 2	100	UN
36	6600335	1 TT + 5 T + 6	5	UN
68	6602947	8 TT + 4 T + 9	"What does notation mean?" E: the way it is written here.	UN
40	6600408	4 TT + 3 T + 5	55	UN
28	6600672	7 TT + 8 T + 1	150	UN
13	6600594	3 TT + 7 T + 9	18	UN
13	6604004	2 TT + 3 T + 6	9	UN
38	6602903	4 TT + 8 T + 2	"Is it plus?" 27.	UN
37	6600188	9 TT + 6 T + 1	"I don't think I know how to do that."	NR
12	6602866	5 TT + 2 T + 7	"I can't get it."	NR
30	6602901	9 TT + 5 T + 8	"I don't know."	NR
64	6600867	2 TT + 7 T + 8	No Response	NR
30	6600437	1 TT + 5 T + 7	No Response	NR
11	6602856	1 TT + 8 T + 5	No Response	NR
62	6600241	4 TT + 8 T + 9	No Response	NR

* Legend: CS = Completely Satisfactory; UN = Unsatisfactory; NR = No Response.

APPENDIX 4, Example 3

Extract from the Report on Results of Testing Unit 16:
Numbers and Measuring, Feb.–Mar.–Apr., 1969*

Item Form 16.11: Addition and Substraction of Half Values

Description: A sheet of notepaper displaying an incomplete number sentence was presented together with

$$\begin{bmatrix}\text{nothing else}\\ \text{a number line}\end{bmatrix}:$$

"Complete this number sentence."

Summary of Correct Responses

Cells 1–16	With Number Line	Without Number Line	Total
a+b (where a and b are drawn from the integers 1 through 9) e.g. 3 + 2.	$\frac{8}{8}$	$\frac{9}{10}$	$\frac{17}{18}$ (94%)
a+b (where a = 1; b = j + ½, with i and j drawn from the integers 1 through 9) e.g., 3 + 2½	$\frac{3}{9}$	$\frac{3}{9}$	$\frac{6}{18}$ (33%)
a+b (where a = i + ½; b = j, with i and j drawn from the integers 1 through 9) e.g., 3½ + 2	$\frac{3}{12}$	$\frac{4}{7}$	$\frac{7}{19}$ (37%)
a+b (where a = 1 + ½; b = j + ½, with i and j drawn from the integers 1 through 9) e.g., 3½ + 2½	$\frac{3}{8}$	$\frac{3}{10}$	$\frac{6}{18}$ (33%)

	With Number Line	Without Number Line	Total
s–b (where s = a + b, with a and b drawn from the integers 1 through 9) e.g. 5–2	$\frac{7}{8}$	$\frac{6}{8}$	$\frac{13}{16}$ (81%)
s–b (where s = a + b, and a = i; b = j + ½, with i and j drawn from the integers 1 through 9) e.g. 5½–2½	$\frac{2}{10}$	$\frac{1}{9}$	$\frac{3}{19}$ (16%)
s–b (where s = a + b, and a = i + ½, b = j, with i and j drawn from the integers 1 through 9) e.g. 5½–2	$\frac{2}{6}$	$\frac{3}{10}$	$\frac{5}{16}$ (31%
s–b (where s = a + b, and a = i + ½, b = j + ½, with i and j drawn from the integers 1 through 9) e.g. 6–2½	$\frac{3}{8}$	$\frac{1}{10}$	$\frac{4}{18}$ (22%)
	$\frac{31}{69}$ (45%)	$\frac{30}{73}$ (41%)	

---

* Prepared by Graham Maxwell.

APPENDIX 4, Example 4

Extract from the Report on Results of Testing Unit 23:
Conditions Affecting Life, Spring 1969*

Item Form 23.10: Explaining how to study effects of moisture on sowbugs or mealworm beetle.**

Cell 1: Categorized Responses (Selected)

Category 1: The responses in this category were the most complete descriptions of how to set up the experiment. All responses mentioned setting up of different conditions of moisture (wet and dry).

"Well I'd take these two and put them in here (picks up blotting paper) then I'd take this and fill it with water (points to petri dish) on one side and let other side dry and see what side mealworm would like best. Then I'd put the lid on and see little air bubbles would come up then the sow bug or mealworm would pop bubble so that they could breathe."

(S says nothing.) (Places blotters in petri dish with space between. Uses eye dropper to put water on 1 piece of blotting paper—carefully wets entire piece. Then places tape on the bottom of dry paper and tapes it to petri dish.)

"Should I set it up?" (Then S puts blotting paper in dish.) "We'd put mealworm in here." (dish) (closes it) (takes eye dropper) "Put water in here (dish) then close the lid—then put the tape over it because otherwise some of the moisture would get out and see how they would react to the moisture on this little container."

Category 2: The responses in this category may have been just as adequate as those in Category 1, but the child's verbalizations were vague.

"We could put these in here (blotting paper in petri dish) put a drop of water in here" (Points to blotting paper and then put a cover over it.)

"You could put these in here (puts blotting paper separated in the petri dish) then you could put a sowbug in here and see where he goes. Then you could water on here." (Points to left half of petri dish.)

"Take a beetle and put it in here (points to dish) and then pour some water in then put two pieces of paper in then put the beetle and see what he does and see if the paper gets wet."

Category 3: The responses in this category constituted incomplete descriptions of how to set up the experiment. None of these responses alluded to setting up contrasting moisture conditions—wet, dry.

"I can't remember how you do that." (S tries to put paper on outside of dish—opens dish—closes it—turns cover upside down—places paper in—takes it out closes it—lays paper on top again only one half.) "Suppose to put a mealworm in there (points to dish) and a piece of paper there (points to top of cover). Don't know how to do it cause we did it—we didn't even need the tape or scissors."

---

* Prepared by Bruce Mussell.

** See Appendix 3 for the complete item form.

"You could dunk those two papers in here (water) then put it in here (dish) then put a mealworm on them."

"Put these together." (Points to all materials.) (How would you put them together?) Put these (blotters) inside dish."

Category 4: The responses in this category referred to some experiment which did not test the effect of moisture on insects.

"Put some mealworm beetles and maybe put some ice in and see if they hibernate or sleep or put cold water in it."

"We can put a sowbug or beetle in here (points to petri dish) then we would put bug in here (points to petri dish) like water cold or warm. I don't know any more."

"We could put them in a cup and put cover on top—that's wrong, take cups like those ones there and fill one with water and put the other on top, then put one of those blue things on top to see how much humidity." (Cups S points to were those making up piece of nature.)

Category 5: No Response

Cell 2: Categorized Responses

Category 1: The responses in this category were the most complete descriptions of how to set up the experiment. All responses mentioned setting up different conditions of moisture (wet and dry).

(RQ) "Put some water on one piece of cloth, and not on the other, then see which one would go on the dry one or the wet one." (No actions.)

Places cloth in container and then removes one piece and dips it entirely into the water then she placed it in plastic container but was not satisfied that it was wet enough, so she redipped it in water and placed it in the plastic container again—leaving a space between pieces of cloth very carefully.

(RQ) "I'm not quite sure. You could put these materials into this plastic thing—then pour a little water, not too much on one side, and see which side they stay on, the wet side or the dry side. You could use the tape to tape it in."

Category 2: The responses in this category may have been just as adequate as those in Category 1, but the child's verbalizations were vague.

"Put a little water on that piece of material. Then put the piece of material on the cover, and see if the mealworm will go off the piece of material."

"Put these down on here . . . sprinkle water on one side, leave the other half dry . . . put the bugs in it." (Points to cloth to put in tray.)

"Take glue or something and put cloth in water and put beetle on and see how he reacts to it." (Student puts cloth in water and places on tray while talking.)

Category 3: Th [illegible]sponses in this category constituted incomplete descriptions of how to set up the experiment. None of these responses alluded to setting up contrasting moisture conditions—wet, dry.

"Hum, well if we put a sowbug or something, we could put it in water to see if he would like it or something." (No actions.)

(RQ "Yes put water on one side of tray with cloth in it and—I don't know what else to do." (A bit puzzled by question.)

"We could put water in what you've got the dirt in now (she was looking at piece of nature on materials tables to my right) and put a sowbug in it and see what he does."

Category 4: The responses in this category referred to some experiment which did not test the effect of moisture on insects.

"I don't really know about this. You get a mealworm and you get a sowbug and you get some water and put them in it and see which one lasts the longest."

(RQ) "Well you'd dig a hole and put the beetle in the hole and see if he'd stay in there or get out."

(RQ) "By fresh air or bad air, or you could try no air."

Category 5: No Response

Data Summary (showing number and percentage of responses in the chosen categories):

	Cell	
Category	1	2
1	15 (50%)	11 (37%)
2	5 (17%)	4 (13%)
3	3 (10%)	7 (24%)
4	6 (20%)	4 (13%)
5	1 (3%)	4 (13%)
Total	30	30

APPENDIX 5: An Example of a General Summary of the Results of Testing a MINNEMAST Unit.

Summary of the Results of Testing Unit 16—Numbers and Measuring (Feb.–Mar.–Apr. 1969)[*]

This summary is an attempt to present in capsule form some of the most important features of the results of the evaluation of Unit 16—Numbers and Measuring. It is not intended to be read without reference to the larger and more detailed analysis of the results forming section 1 of the total report. However, it should provide the reader with an overview of the results and a means for discovering what aspects of the more complete analysis he may wish to explore further. In addition, it will raise questions of the significance of the results, especially in relation to the general aims of the unit and of the whole MINNEMAST Curriculum, so that further discussion of the implications of these results may be encouraged. The treatment moves generally through the item forms in numerical order. It is hoped that this procedure will facilitate reference to section 1 of the report.

The test required the administration of 2743 test items to a total of 396 Grade 2 children from 16 classes in 5 MINNEMAST schools in the Twin Cities area. All children in this experimental population had been taught the relevant MINNEMAST Unit (16—Numbers and Measuring) during the Fall and Winter of the 1968–69 school year. Testing was done on an individual basis by three experimenters and was begun as soon as possible after the teaching of the lesson had been completed. All testing was done during the months of February, March, and April 1969. No child was administered more than 5 items for each part of the total test, i.e., 15 items for all three parts. All three parts of the test were administered at the same sitting. Details of the number of items administered are shown in Table 1.

Table 1: Distribution of Items Among Children

Part[*]	0 items	Number of children receiving 1 item	2 items	3 items	4 items	5 items	Total Items	Av/Child
A (16.1–16.11)	23	79	91	85	60	58	1046	2.64
B (16.12–16.20)	43	105	104	83	36	25	831	2.09
C (16.21–16.29)	46	84	119	72	47	28	866	2.18
						Totals	2743	6.92

[*] The three parts constituted three separate tests administered consecutively. Part A encompassed sections of the unit dealing with Arithmetic, Part B with Measuring Weight, and Part C with Measuring Length.

Item Form 16.1 required children to name a three-digit numeral. Percentages of satisfactory responses were 84% for the even hundreds, 79% for numerals with a zero in the middle (tens) positions, 88% for numerals with a zero in the units position, and 100% for numerals with no zero present. The results indicate most children had relatively little difficulty in naming three-digit numerals. As expected, most difficulty occurred for numerals with a zero in the middle position.

---

[*] Prepared by Graham Maxwell.

Egyptian Numeration was not stressed in the unit (being dealt with in a single lesson and three worksheets) so it is not surprising that performance on Item Form 16.2 is poor. Only 22% of the children could correctly represent a three-digit Arabic numeral in the Egyptian Numeration system. For the opposite task, representing an Egyptian numeral as a three-digit Arabic numeral, 32% of the children produced a correct response when the Egyptian symbols were in descending order of magnitude, but only 17% could do so when the symbols were ordered randomly. A few more children (22%, 12%, and 17% respectively) made responses that involved small errors in counting or choice of symbols, but the majority (56%, 56%, and 66% respectively) produced responses that seem to indicate more confusion than understanding. If the lesson is considered important enough for inclusion in the unit, then sufficient practice to remove the present confusion would seem desirable.

Roman Numeration also received attention in only one lesson and three worksheets. Item Form 16.3 focused on specific numbers in the range 1 through 21. Cell frequencies are too low (5 per cell) to enable meaningful analysis of the results for each specific number. The percentage of correct responses for representing Arabic numerals as Roman numerals was 33% and for representing Roman numerals as Arabic numerials was 50%. Although these figures indicate that a majority of the children did not learn the simplest aspects of the Roman Numeration system, they are nevertheless encouraging. For one thing, instruction was brief, and for another the item form does not test all that may have been learned (symbols L and C were also used in instruction). Obviously, the opportunity for revision and further practice at some stage in the curriculum would be desirable.

Item Forms 16.4 and 16.5 deal with Decimal Place-Value Notation. The tasks involved the use of a set of little squares, strips of ten little squares, and cards of ten strips. The relations among the objects were stated but the fact that each card contained one hundred little squares was not. One task was to choose objects from the total set so that the total number of little squares chosen was that indicated by a three-digit numeral. The second task was the converse of the first—to write a three-digit numeral showing the total number of little squares in a given set. The results are shown in Table 2.

Table 2: Percentages of Completely Satisfactory Responses for Item Forms 16.4 and 16.5

Numeral	I.F. 16.4 Given Numeral, Produce Set	I.F. 16.5 Given Set, Produce Numeral
i0k	16%	16%
ij0	44%	28%
ijk	20%	24%

Obviously, the results do not support the contention that place-value is well understood. The tasks themselves were relatively simple and represent a straightforward generalization from those in the unit. Difficulties did arise for some children from the form of the instructions and the nature of the materials. But these considerations do not account for many of the incorrect responses. Obviously, any future application of these item forms at this level should include cells involving two-digit numerals, cells with various cues to the place-value system (such as pointing out that each card contains 100 little squares), and possibly cells with different materials (such as pellets in transparent packets and bottles or counters in open cups and trays as in the unit).

The use of T-Notation to represent numerals is closely linked to the understanding of place-value. Item Form 16.6* was designed to investigate both the representational and the place-value aspects of T-notation. The results were the worst for any item form in this test and are shown in Table 3.

Table 3: Percentages of Completely Satisfactory Responses for Item Form 16.6

Description of Cells	Percent Correct
1. i T.T + 0 T + k = ___	12%
2. i T.T + k = ___	4%
3. i T.T + j T + 0 = ___	16%
4. i T.T + j T = ___	4%
5. i T.T + j T + k = ___	4%
6. ijk = ___	0%
7. ijk = ___ T.T + ___ T + ___	40%

(where for any item i, j, and k are non-equal digits from the range 1 through 9.)

It is quite clear that the great majority of the children could not use or interpret the symbolization of three-digit numerals in T-notation. This is serious since T-notation is used as one of the central teaching strategies of the curriculum (for example, for teaching about place-value, number base, algebraic symbolization, power and indices). There is no comfort in the percentage (40%) obtained in the seventh cell, since not only does it still represent a minority of the children with the remainder giving either largely nonsensical responses (28%) or no responses at all (32%), but also the evidence from other cells of the item form is that these responses represent simply rote memorization of a procedure where the three digits of the numeral were placed consecutively in the blank spaces provided. Almost no one could do the reverse operation (although it was dealt with in the worksheets) (c.f. Cells 1, 3, and 5) and little to no account was taken of the symbol that signifies the values of each digit (c.f. Cells 1 and 3 and compare them respectively with Cells 2 and 4). Even more worrisome is the discovery that 17% of all responses in Cells 1 though 5 seem to have ignored the T symbols entirely and made a simple addition of the digits, and that a further 44% were not only wrong but nonsensical.

Taking Item Forms 16.4, 16.5, and 16.6 together, the evidence is clear that place-value was not well understood by these children.

It has been claimed (Unit 16, p. 75) that the Base-Four Number System was taught "only to help the children achieve a better understanding of place-value. It is not intended that they should master the concepts and ideas of the base-four numeration system. The purpose of having the children work with the base-four system, in which they group by fours, is to have them comprehend the place-value concept more easily as it applies to the decimal system, in which they group by tens." The exact nature of what will be learned about base-four numeration and the mechanism by which this learning will transfer to base-ten numeration is not stated. Presumably, something must be learned for transfer to occur. On the basis of the content of the three lessons devoted to this topic (Unit 16, pp. 75–86), it can be assumed that such learning must at least include understanding the terms "base 4" and "base 10" and the idea of grouping into 4's and 10's according to the specified base. These understandings could be demonstrated by correct representation in base 4 and base 10, especially the latter, of the number of objects in a set, and also by construction of a set of objects whose number is specified in base 4 or base 10, especially the latter. These are the tasks set by Item Forms 16.7 and 16.8. In each cell the question was asked with respect to both base 4 and base 10, with order varied between cells. The results show that only one child in 24, i.e., 4% could produce the correct response in base 4. These same children could also produce the correct response in base 10.

---

* See Appendix 4, Example 2.

Since the unit claims that acquaintance with base 4 will facilitate understanding of base 10, it is interesting to look more closely at the parts relating to base 10 in the responses to each item. Table 4 summarizes the results for these partial responses.

Table 4: Percentages of Correct Responses to that part of each item in Item Forms 16.7 and 16.8 requiring use of Base-10 Numeration.

	Base 4, then Base 10	Base 10, then Base 4
Producing number in Base 10	54%	67%
Producing set in Base 10	44%	62%

The results are consistent with the interpretation that questions about base 4 inhibit rather than enhance performance in base 10. Moreover, even when questions about base 10 were presented first, the percentages of correct responses were not very high. The question needing to be answered is whether the hope that (unspecified) features of the base-10 number system will be learned by transfer from an incompletely learned base-4 number system is psychologically and pedagogically sound. The evidence of these results suggests that it is not.

The results for Item Form 16.9 in which use was made of the abacus are difficult to analyze. The items were not completely successful in eliciting the desired kinds of responses since the instructions did not make it unambiguously clear that the second number was to be added to or substracted from the first number. Thus some children simply set up the abacus to represent each number in turn. It is probably most useful to examine the percentage of responses which, quite apart from this misinterpretation of the instructions and quite apart from simple counting errors, must be scored incorrect. In fact, an average of 40% of the children (36% for Cell 1 and 44% for Cell 2) produced responses that demonstrate incomplete or faulty understanding of the abacus and hence also of place-value notation. The nature of these incorrect responses may be suggestive of ways in which instruction in the use of the abacus and of place-value notation could be improved.

Item Form 16.10 was designed to test addition and substraction of small positive and negative numbers with the aid of the Number Line or the Slide Rule. The relevant parts of Unit 16 are: Lesson 7—The Number Line; Lesson 9—The Abacus; Lessons 26 and 27—Directions on the Number Line and Negative Integers. Cells involving use of the Slide Rule for problems in subtraction and in the use of negative integers were included as interesting generalizations of the content of the curriculum. In the event, certain of the cells in the item form were omitted as being too extreme. The results for those cells that were included are given in Table 5.

Table 5: Combined Percentages of Completely Satisfactory and Partially Satisfactory Responses for Item Form 16.10

	Number Line	Slide Rule	Totals
$n_1 + n_2$ (positive integers)	8/9 (89%)	8/11 (73%)	16/20 (80%)
$n_1 - n_2$ (→ pos. integer)	5/9 (56%)	6/11 (55%)	11/20 (55%)
$n_1 - n_2$ (→ neg. integer)	2/8 (25%)	3/12 (25%)	5/20 (25%)

The results provide no evidence that the Number Line aided computation any better than the Slide Rule. This is somewhat surprising—since "subtraction" and "negative integers" were taught using a Number Line but not a Slide Rule. On the other hand, there are large differences among the row totals. The conclusion must be that "subtraction" and particularly "subtraction resulting in a negative integer" has been poorly learned. Further evidence of the irrelevance of the number line or slide rule in the performance of these computations is seen in the number of correct responses made without using these aids at all and sometimes made in spite of using the aids incorrectly.

Item Form 16.11* was designed to test competence in the addition of half-values with the aid of a number line (Cells 1–4). This was specifically taught by the unit. Generalizations were also investigated—addition of half-values without the aid of a number line, and subtraction of half-values with and without the number line. It is surprising to note that the presence or absence of the number line appears to have had no influence on the children's performance, and that subtraction was performed only slightly worse than addition. The addition and subtraction of two integers was in all cases performed with very little difficulty (88% correct responses). However, the primary focus of the item form, viz, addition of two numbers with the aid of the number line where at least one number carries the fraction "½", showed consistently poor results (31% correct responses). It is perhaps not surprising that performance was so poor given the extremely brief attention given to it in the unit.

The percentage of children who can read a symbolic comparison statement satisfactorily is at best 62% ($L_A = L_B$) and at worst 38% ($W_a < W_b$ and $W_a > W_b$). The two item forms requiring translation of comparison statements are 16.12 (weight) and 16.21 (length). Except for the obvious difference in the property symbols, the only other difference between the two item forms was that the

Table 6: Combined Percentages of Completely Satisfactory and Partially Satisfactory Responses for Item Forms 16.12 and 16.21

	Weight	Length
<	38%	42%
>	38%	58%
=	50%	62%
Combined	43%	54%

---

* See Appendix 4, Example 3.

subscripts for W were lower-case while the subscripts for L were upper-case. The cell percentages obtained for generally satisfactory responses are shown in Table 6.

There seems to have been slightly better performance with length symbols than with weight symbols. It is difficult to account for this difference unless it is due to more practice in making length comparisons and making length comparison statements, especially as a result of Unit 12.

Obviously, later instruction should not be designed on the assumption that almost all the children who have been taught Unit 16 will be able to specify the verbal equivalent of written symbolic comparison sentences. Only about one-half of the children can, in fact, do this satisfactorily. In defense of Unit 16, we can point to the novelty of the full-blown symbolism (in the previous unit on measurement, Unit 12, property and object identifications were made in words, e.g., Length of A < Length of B) and the relatively small amount of instruction and practice—Lessons 1, 2, and 19—given to it in Unit 16. Against that must be contrasted the fact that most of the responses classified as incorrect were grossly wrong and involved stating the relation incorrectly, not stating the relation at all, or making no relevant response.

Item Forms 16.13 and 16.14* required the beam balance to be used for comparing the weights of two objects. In one case responses were verbal in the other written. The percentages of satisfactory responses were uniformly high and are independent of the relative size of the objects. Table 7 shows a summary of the results.

Table 7: Percentages of Completely Satisfactory and Partially Satisfactory Responses Pooled Across Size Variation for I.F. 16.13 and I.F. 16.14

	$W_l > W_r$	$W_l < W_r$	$W_l = W_r$
Making Verbal Statement of Relation	70%	73%	75%
Completing Written Statement of Relation	70%	62%	78%

Whereas the ordering of two objects by weight proved to have been well learned, the children were much less successful in ordering three or four objects. The percentages of satisfactory responses in Item Form 16.15 were 32% for ordering three objects, and 21% for ordering four objects. Most of the unsatisfactory responses were not even very adequate attempts at performing the task.

Surprisingly, weighing an object on the beam balance was a task most children could not perform. In Item Form 16.16, by combining across all cells (different weighing Units), we discover that only 13% of the children were able to carry out the correct weighing operation, and only half of those could report the result of their weighing. Even more distressing is the revelation that 62% of the children did nothing but put the object into the balance-cups. There would seem to be an obvious weakness in the instruction at this point and it seems likely that the majority of children were never given the opportunity of manipulating the balance themselves. Since the assumption of skill in weighing is a prerequisite of some later units (notably Units 19 and 24), more opportunity for acquiring this skill is probably desirable.

Item Form 16.17 was an attempt to investigate whether the children could establish the weight relation between two objects without making a direct comparison of them. At least two methods of performing the task exist: (i) weigh each object on the balance and compare the two weight values; (ii) balance one object against a set of standard weights and then compare the second object with the same set. The principle of transitivity is involved in both methods. Unfortunately, since few children were able to use the beam balance correctly (cf. Item Form 16.16), this task proved too difficult for most children. Only four children out of sixty-four (6%) could produce the kind of response asked for. Another three children (5%) went as far as weighing one of the objects but stopped at that point, and 10 more (16%) made a direct comparison and produced the correct relation.

---

* See Appendix 3 for Item Form 16.14.

The beam balance requires equal arms for accurate weighing. Changing the length of the arms or the distance of a suspended object from the fulcrum does not change the weight but will change the attitude of the balance. Item Form 16.18 investigated what was known about these aspects of the functioning of the beam balance. Because of the sensitivity of the balance and the complexity of the questions, the results are not easy to summarize, but it does seem that at least half the children do not understand the significance of having equal vs. unequal arms. When the arms were initially unequal and level, 53% of the children said that the objects were equal in weight. A little over a third of the chidren (37%) could make a satisfactory prediction concerning the balance for changing from unequal to equal arms, although a larger number (53%) could do so for changing from equal to unequal arms. However, even in some of these responses there was a verbal confusion between the property of an object called "weight" or "heaviness" and the attitude of the beam balance which signifies the relation between two objects for that property. To some children, when the attitude of the beam balance changes (here by altering arm-lengths), the weights of the objects change.

Item Form 16.19 was also concerned with the equality of the arms of the beam balance except that here the focus was on the symmetry effects of this equality. Once again the complexity of the responses cannot be captured in a brief summary. However, the results demonstrate that the majority of the children (77% with equal-weight objects; 85% with unequal-weight objects) could make a satisfactory prediction of the effect of interchanging the objects. Once again, the verbal description given by some children left their intention regarding conservation of weight as against change in the attitude (if any) of the balance unresolved.

The use of a spring balance to compare the weights of two objects (Item Form 16.20) was successfully demonstrated by 81% of the children. These were easily the best-performed of all the items on weight and it is to be wondered that more use has not been made of this simple device in teaching about weight and weight measurement.

Item Form 16.21 has already been discussed with Item Form 16.12. The comparison of lengths with production of a verbal statement of the relation (Item Form 16.22) was successfully performed by about 80% of the children. Discriminations were made with equal ease whether the difference in length was ⅛", ⅜" or 2". However, technical language was almost non-existent in the responses—for example, no child gave the form of response "The length of A appears to be the same as the length of B," and gave mostly responses of the form "They're both the same" or "This one's larger."

Item Form 16.23 was similar to the previous one except that a response was made by inserting one of the three comparison symbols in an incomplete symbolic sentence. When the objects were as near as possible equal in length, 74% correctly made the comparison and wrote =. When the lengths were unequal, 60% to 68% correctly made the comparison and wrote the relevant < or > except in the case of inserting > for objects ⅛" different in length (33%). This last result may be explainable in terms of sampling error caused by having to exclude certain items which were incorrectly presented.

Item Form 16.24, which required the placing of several objects in order by length, also produced high percentages of correct responses; 87% for three objects; 88% for four objects; 75% for five objects.

Item Form 16.25 involved measuring the length of an object and reporting that length to a specified approximation. The results are shown in Table 8.

Table 8: Combined Percentages of Completely Satisfactory and Partially Satisfactory Responses for Item Form 16.25

	Actual Length Larger	Actual Length Smaller
Approx. to nearest inch	50%	15%
Approx. to nearest half-inch	45%	40%
Approx. to nearest cm.	45%	30%

Errors were mostly made in choosing the correct scale or making the correct approximation. (The top two percentages in Table 8 could be boosted to 65% and 60% if nearest-half-inch responses were counted correct.) There are few responses (16% of the total) that demonstrate complete incompetence in measuring.

Item Form 16.26 parallels the previous one. Here the child marks his response on a number line. The percentages of satisfactory responses are of the same order as for verbal response. They are given in Table 9.

Table 9: Combined Percentages of Completely Satisfactory and Partially Satisfactory Responses for Item Form 16.26

	Actual Length Larger	Actual Length Smaller
Approx. to nearest inch	40%	20%
Approx. to nearest half-inch	50%	20%
Approx. to nearest cm.	50%	35%

The results of both this and the previous item form point fairly consistently to the conclusion that making an approximation to the next largest unit is more troublesome than making an approximation to the next lowest unit. The children experienced more difficulties in ruling a line of specified length (Item Form 16.27) than were anticipated. Some are an artifact of the presentation (a few children did not alter the relative positions of response sheet and ruler). It is interesting that ruling whole inches or whole centimeters should produce percentages of satisfactory responses (55% and 50% respectively) consistent with the previous item forms on measuring length, but that ruling a fractional number of inches (here 5½ inches) should prove inexplicably troublesome (only 25% satisfactory responses).

Item Form 16.28 required comparison of two lines on opposite sides of a card using different techniques. Results are shown in Table 10.

Table 10: Combined Percentages of Completely Satisfactory and Partially Satisfactory Responses for Item Form 16.28

	Ruler	Rod and Pencil	Set of Minnebars
Equal Lengths (7½")	67%	50%	92%
Unequal Lengths (7½" and 8")	67%	75%	39%

The apparent irregularities in these data are intriguing. Tentative explanations can be attempted. When using a rod and pencil, small imprecisions will lead to equal lengths being pronounced unequal, whereas judgments about unequal lengths will be unaffected by small imprecisions. When using a set of minnebars, comparing equal lengths would appear to involve fewer choices concerning the next operation than comparing unequal lengths; at any rate, the actual responses were simpler for the former than the latter.

In Item Form 16.29 a set of symbolic statements about the lengths of some hypothetical objects were presented. The task was to fill in the correct letters for the objects in the blank spaces of an incomplete symbolic comparison sentence. With information about two objects only, 30% of the children completed the comparison sentence by filling in the correct letters; another 20% using the numbers (of inches) instead of object letters. A much smaller percentage of the children could produce either kind of response for 3 objects (20%), 4 objects (10%), and 5 objects (25%). Obviously, they were not yet at home with the use of symbolic notation.

On the whole, the results indicate better understanding and skill in length measurement than in weight measurement. This was to be expected since length and length measurement have been treated in previous units. Obviously, the unit has made some progress toward teaching the central concepts and skills embodied in it. However, the results presented here also remind that it can by no means be assumed that all, or even in some cases most, children have learned the central concepts and skills of number place-value, weight measurement, and length measurement. At best, a beginning has been made and much practice and review will be necessary to expand and consolidate that beginning. Further, the results pose a warning that there are some cases of untoward generalizations that may present problems of unlearning at later stages in the curriculum.

REFERENCES

Alkin, M. C. Evaluation theory development. *Evaluation Comment*, 1969, 2(1), 2–7.

Alkin, M. C. Evaluating the cost-effectiveness of instructional programs. In M. C. Wittrock & D. E. Wiley (Eds.), *The evaluation of instruction*. New York: Holt, Rinehart, and Winston, 1970.

Bloom, B. S. Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation, new roles, new means*. (68th NSSE Yearbook) Chicago: University of Chicago Press, 1969.

Bormuth, J. R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Brickell, H. M. Appraising the effects of innovations in local schools. In R. W. Tyler (Ed.), *Educational evaluation, new roles, new means*. (68th NSSE Yearbook) Chicago: University of Chicago Press, 1969.

Center for the Study of Public Policy. *Educational vouchers: A report on financing elementary education by grants to parents.* Cambridge, Massachusetts: U. S. Office of Economic Opportunity, Office of Program Development, 1970.

Chase, F. S. Educational research and development in the sixties: The mixed report card. Background paper submitted to the Select Subcommittee on Education, U. S. House of Representatives, Washington, 1971.

Cronbach, L. J. Course improvement through evaluation. *Teachers College Record,* 1963, *64,* 672–683.

Gagné, R. M. Curriculum research and the promotion of learning. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation,* AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, *18,* 519–521. (Reprinted in W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction.* Englewood Cliffs, N.J.: Educational Technology Publications, 1971.)

Glaser, R. A criterion-referenced test. In W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction.* Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Gottfried, N. W., & Ryan, J. J. Arithmetic achievement test performance of MINNEMAST mathematics pupils in the third and fourth grades. Longitudinal Assessment of the MINNEMAST Mathematics Program, Report No. 1. MINNEMAST Project, University of Minnesota, 1968.

Hemphill, J. K. Educational development. *Urban Review,* 1969, 23, 23–27.

Hively, W. A test-item pool for MINNEMAST science unit 2.1: Measuring weight. MINNEMAST Project, University of Minnesota, 1966.

Hively, W. Introduction to domain-referenced achievement testing. Paper presented at Annual Convention of the American Educational Research Association, March, 1970. MINNEMAST Project, University of Minnesota, 1970.

Hively, W., Patterson, H. L., & Page, S. A universe defined system of arithmetic achievement tests. *Journal of Educational Measurement,* 1968, 5, 275–290.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., & Patalino, M. *CSE elementary school test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Hoepfner, R., Stern, C., Nummedal, S. G., Doherty, W. J., DeMuth, J., Marshall, J., Simon, L., & Gaffney, J. *CSE-ECRC preschool/kindergarten test evaluations.* Los Angeles: Center for the Study of Evaluation, University of California, 1971.

Hoepfner, R., Hemenway, J., DeMuth, J., Tenopyr, M. L., Granville, A. C., Petrosko, J. M., Krakower, J., Silberstein, R., Nadeau, M-A., Boyer, E. G., Hill, R. A., Levin, J., Nickleach, S., Thomas, J., & Simon, A. *CSE-HLP test evaluations: Tests for higher-order cognitive, affective, and interpersonal skills.* Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Hoepfner, R., Bradley, P. A., Klein, S. P., & Alkin, M. C. *CSE elementary school evaluation kit: Needs assessment.* Boston: Allyn and Bacon, 1973. (in press)

Husek, T. R., & Sirotnik, K. Matrix sampling. *Evaluation Comment,* 1968, *1*(3), 1–4.

Klein, S. P., Burry, J., Churchman, D., & Nadeau, M-A. *Evaluation workshop I: An orientation*. Monterey, California: CTB/McGraw Hill, 1971.

Klein, S. P., Burry, J., & Churchman, D. *Evaluation workshop II: Needs assessment*. Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Lindvall, C. M., & Cox, R. C *Evaluation as a tool in curriculum development: The IPI program*. AERA Monograph Series on Curriculum Evaluation, No. 5. Chicago: Rand McNally, 1970.

Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, 72, 143–155.

Mager, R. F. *Preparing instructional objectives*. Palo Alto, California: Fearon, 1962.

McNeil, J. D. *Toward accountable teachers*. New York: Holt, Rinehart, and Winston, 1971.

Merwin, J. C., & Womer, F. B. Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In R. W. Tyler (Ed.), *Educational evaluation, new roles, new means*. (68th NSSE Yearbook) Chicago: University of Chicago Press, 1969.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, 28, 95–104.

Platt, J. *How men can shape their culture*. University of Michigan, Mental Health Research Institute Communication 282, 1970.

Popham, W. J. (Ed.) *Criterion-referenced measurement: An introduction*. Englewood Cliffs, N.J.: Educational Technology Publications, 1971. (a)

Popham, W. J. Performance tests of teaching proficiency: Rationale, development and validation. *American Educational Research Journal*, 1971, 8(1), 105–117. (b)

Popham, W. J. *An evaluation guidebook: A set of practical guidelines for the educational evaluator*. Los Angeles: The Instructional Objectives Exchange, 1971. (c)

Popham, W. J., & Baker, E. L. *Systematic instruction*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, 6(1), 1–9. (Reprinted in W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction*: Englewood Cliffs, N. J.: Educational Technology Publications, 1971.)

Rabehl, G. The experimental analysis of educational objectives. Unpublished doctoral dissertation, University of Minnesota, 1971.

Rosenbloom, P. C. Science and mathematics in the curriculum. Invited Address, National Council of Teachers of Mathematics. MINNEMATH Center, University of Minnesota, 1964.

Rosenbloom, P. C. The relations between mathematics and the elementary science curriculum. Address to Minnesota Education Association and Minnesota Federation of Teachers. MINNEMATH Center, University of Minnesota, no date. (a)

Rosenbloom, P. C. A brief overview of the MINNEMAST mathematics program. MINNEMATH Center, University of Minnesota, no date. (b)

Schutz, R. E. The nature of educational development. *Journal of Research and Development in Education*, 1970, 3(2), 39–64.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Sension, D. A comparison of two conceptual frameworks for teaching the basic concepts of rational numbers. Unpublished doctoral dissertation, University of Minnesota, 1971.

Skager, R. System for objectives-based evaluation—Reading. *Evaluation Comment,* 1971, *3*(1), 6–11.

Smith, E. R., Tyler, R. W., et al. *Appraising and recording student progress.* Vol. 111 of the Adventure in American Education Series describing the Eight Year Study. New York: Harper and Brothers, 1942.

Stake, R. E. The countenance of educational evaluation. *Teachers College Record,* 1967, *68*(7), 523–540. (a)

Stake, R. E. Toward a technology for the evaluation of educational programs. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation.* AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967. (b)

Stake, R. E. National assessment. Proceedings of 1970 Invitational Conference on Testing Problems. Princeton: ETS, 1970.

Stake, R. E., & Gooler, D. Measuring educational priorities. *Educational Technology,* 1971, *11*(9), 44–49.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. *Educational evaluation and decision making.* Itasca, Illinois: Peacock Publishers, 1971.

Tyler, R. W. *Constructing achievement tests.* Columbus, Ohio: Bureau of Educational Research, 1934.

Tyler, R. W. Purposes and procedures of the evaluation staff. In E. R. Smith, R. W. Tyler, et al. *Appraising and recording student progress.* The Eight Year Study. New York: Harper and Brothers, 1942.

PS73-205/5M/5-73